Lehrstuhl für Bioinformatik
Fakultät für Informatik
Technische Universität München

Data Mining Methods for Medical Diagnosis

Test Selection, Subgroup Discovery, and Constrained Clustering

Marianne Larissa Mueller

Vollständiger Abdruck der von der Fakultät für Informatik
der Technischen Universität München
zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Rüdiger Westermann
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Stefan Kramer, Johannes Gutenberg Universität Mainz
	2.	Univ.-Prof. Dr. Frank Puppe, Julius-Maximilans-Universität Würzburg

Die Dissertation wurde am 21.06.2011 bei der Technischen Universität München eingereicht
und durch die Fakultät für Informatik am 20.03.2012 angenommen.

Abstract

As in many areas of life also in medical domains, the volume and complexity of data collected and stored is growing at a rapid pace. In the medical field, this includes besides information obtained in clinical studies also secondary data from patient records, images, and test results. Even though this information is valuable for diagnosis and therapy, it is rarely ever used after the actual treatment and the release of the patient. Based on a more comprehensive analysis of such data, it may be possible to both come up with new medical hypotheses and find statistical evidence for existing hypotheses. This approach should partially overcome the limitations of traditional medical studies and trials with a fixed hypothesis to be confirmed or refuted, and a small number of subjects.

In this thesis, we propose new techniques for *test selection*, *subgroup discovery*, and *constrained clustering* for the improved diagnosis in two medical domains, breast cancer and Alzheimer's disease. For the domain of breast cancer diagnosis, we mainly focus on the task of test selection. For the domain of Alzheimer's disease we focus on the correlation of image and non-image data.

We propose two approaches to test selection, one based on information maximization and one on subgroup discovery. In the first approach, we use the concept of *conditional mutual information (CMI)* to select test variables for medical diagnosis. Computing CMI requires estimates of joint distributions over collections of background variables. However, computing accurate joint distributions conditioned on a large set of variables is expensive in terms of data and computing power. Therefore, we propose a data-efficient approach that enables conditioning on all background variables at once by making some conditional independence assumptions. In the second approach, test selection is based on the discovery of subgroups of patients sharing the same optimal test. Subgroups are defined in terms of background information about the patient. We present an algorithm that determines the top subgroups a patient belongs to, and decide for the test proposed by their majority. In experiments, we show that significant parts of the search space can be pruned due to favorable properties of the scoring function.

While we developed a subgroup discovery algorithm for test selection for the breast cancer domain, we investigated its use in the context of clustering and correlating image and non-image data in the dementia domain. Here, the goal is to start with the images and explain the differences and commonalities in terms of the other non-image variables. After clustering PET scans of patients to form groups sharing similar features in brain metabolism, we explain the clusters by relating them to non-image variables, using an algorithm for relational subgroup discovery. Experimental results hint at differences in brain metabolism in terms of demographic and clinical variables. Another possible approach to correlate image and non-image data is to apply methods for constraint-based clustering. In contrast to the previous approach, in which image clusters were

determined without constraints and subsequently explained by non-image data, here the image clusters are constrained by conditions on non-image data in the first place. Frequent itemsets on non-image data describe candidate clusters, which still have to be combined into an overall clustering. This combination can be controlled by a variety of user-defined constraints. The constraints may concern the type of clustering (e.g., complete clusterings, overlapping or encompassing clusters) and the composition of clusterings (e.g., certain clusters excluding others). We show that these constraints can be translated into integer linear programs, which can be solved by standard optimization packages. In our experiments with dementia data we show the trade-offs involved in balancing various kinds of constraints.

Zusammenfassung

Wie in den meisten Lebensbereichen nehmen auch im medizinischen Umfeld das Volumen und die Komplexität der gesammelten und gespeicherten Daten immer schneller zu. Neben Daten, die in klinischen Studien explizit erhoben werden, gibt es eine Fülle medizinischer Sekundärdaten aus Patientenakten, Bildern und Testergebnissen. Obwohl diese Daten wertvolle Informationen für Diagnose und Therapie bergen, werden sie nur selten auch nach der Behandlung und Entlassung des Patienten eingesetzt. Eine umfassendere Analyse dieser Daten kann dazu beitragen, sowohl neue medizinische Hypothesen aufzustellen, als auch vorhandene Hypothesen statistisch zu belegen. Mit diesem Ansatz kann man die Defizite herkömmlicher medizinischer Studien überwinden, in denen eine feste Hypothese anhand einer kleinen Versuchspersonenzahl bestätigt oder widerlegt wird.

In der vorliegenden Dissertation stellen wir neue Techniken für *Test Selection*, *Subgroup Discovery* und *Constrained Clustering* vor, um die Diagnose in zwei medizinischen Domänen, Brustkrebs und Alzheimer, zu verbessern. Im Bereich Brustkrebsdiagnose liegt unser Schwerpunkt auf *Test Selection*. Im Bereich Alzheimer besteht unser Hauptinteresse in der Korrelation von Bilddaten und strukturierten Daten.

Wir stellen zwei Methoden zur *Test Selection* vor. Eine basiert auf der Maximierung von Information, die andere auf Subroup Discovery. Im ersten Ansatz verwenden wir das Konzept der *Conditional Mutual Information (CMI)* zur Auswahl der geeigneten Testvariablen. Zur Berechnung der CMI muss man die Verbundwahrscheinlichkeiten über alle Hintergrundvariablen abschätzen, was bei einer hohen Zahl von Hintergrundvariablen teuer hinsichtlich Daten und Rechenleistung ist. Deshalb stellen wir einen dateneffizienten Ansatz vor, der alle Hintergrundvariablen einbeziehen kann, indem einige bedingte Unabhängigkeitsannahmen getroffen werden. Der zweite Ansatz zur *Test Selection* basiert auf der Entdeckung von Untergruppen von Patienten, die denselben optimalen Test haben. Untergruppen sind dabei über Hintergrundinformationen der Patienten definiert. Wir präsentieren einen Algorithmus, der die besten Untergruppen, in die ein Patient gehört, bestimmt und sich dann für den Test entscheidet, der von der Mehrheit der Gruppen vorgeschlagen wird. In Experimenten zeigen wir, dass grosse Teile des Suchraums aufgrund günstiger Eigenschaften der Scoring Funktion ausgelassen werden können. Nachdem wir einen *Subgroup Discovery* Algorithmus zur *Test Selection* in der Brustkrebs Domäne präsentieren, untersuchen wir seine Verwendung im Kontext von Clustering und der Korrelation von Bilddaten und strukturierten Daten von Demenzpatienten. Hier ist das Ziel, von den Bildern auszugehen und dann die Unterschiede und Gemeinsamkeiten mit anderen Daten zu erklären. Zunächst clustern wir PET Scans, um Gruppen von Patienten mit ähnlichen Gehirnaktivitätseigenschaften zu erhalten. Dann beschreiben wir die Cluster durch strukturierte Daten mit den Ergebnissen eines Algo-

rithmus für *Relational Subgroup Discovery*. Unsere experimentellen Ergebnisse weisen darauf hin, dass es Unterschiede in der Gehirnaktivität gibt hinsichtlich der demographischen und klinischen Variablen.

Eine weitere Möglichkeit, um Bilddaten und strukturierte Daten zu korrelieren, ist der Einsatz von Methoden des *Constrained Clusterings*. Im Gegensatz zum vorherigen Ansatz, in welchem Bilder ohne Vorbedingungen geclustert und anschließend durch strukturierte Daten erklärt wurden, wird hier das Clustern der Bilder durch Vorbedingungen auf den strukturierten Daten eingeschränkt. *Frequent Itemsets* beschreiben die Kandidatencluster, die noch zu einem Gesamtclustering kombiniert werden müssen. Diese Kombination kann durch eine Vielzahl von benutzerdefinierten Bedingungen gesteuert werden. Die Bedingungen können das Clustering und die Konstruktion des Clusterings betreffen. Wir zeigen, dass man diese Bedingungen in ganzzahlige lineare Programme übersetzen kann, welche durch Standardoptimierungspakete gelöst werden können. In unseren Experimenten auf den Demenzdaten zeigen wir, wie die verschiedenen Bedingungen das Ergebnis steuern.

Publications

Parts of this thesis have been pre-published:

Conferences

- Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Barat Rao, and Stefan Kramer. Subgroup Discovery for Test Selection: A Novel Approach and its Application to Breast Cancer Diagnosis. In: *Advances in Intelligent Data Analysis VIII*, 8th International Symposium on Intelligent Data Analysis, IDA 2009, ed. by Niall M. Adams, Céline Robardet, Arno Siebes, Jean-François Boulicaut, vol. 5772, pp. 119-130, Springer. Lecture Notes in Computer Science, 2009.

- Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Barat Rao, and Stefan Kramer. Data-Efficient Information-Theoretic Test Selection. In: *Proceedings of the 12th Conference on Artificial Intelligence in Medicine*, ed. by Carlo Combi, Yuval Shahar, Ameen Abu-Hanna , vol. 5651, pp. 410–415, Springer. Lecture Notes in Computer Science, 2009.

- Andreas Hapfelmeier, Jana Schmidt, Marianne Mueller, Robert Pernetzky, Alexander Drzezga, Alexander Kurz, and Stefan Kramer. Interpreting PET Scans by Structured Patient Data: A Data Mining Case Study in Dementia Research In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, pp. 213–222, 2008.

- Marianne Mueller and Stefan Kramer. Integer Linear Programming Models for Constrained Clustering. In: *Proceedings of the 13th International Discovery Science (DS 2010)*, ed. by Bernhard Pfahringer, Geoffrey Holmes and Achim Hoffman, vol. 6332, pp. 159–173, Springer. Lecture Notes in Computer Science, 2010.

Journals

- Jana Schmidt, Andreas Hapfelmeier, Marianne Mueller, Robert Pernetzky, Alexander Drzezga, Alexander Kurz, and Stefan Kramer. Interpreting PET Scans by Structured Patient Data: A Data Mining Case Study in Dementia Research In: *Journal of Knowledge and Information Systems (KAIS)*, vol 24, pp. 149–170, 2010.

Technical Reports

- Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Barat Rao, and Stefan Kramer. *Data-Efficient Information-Theoretic Test Selection* TU München, Technical Report TUM-I0910, 2009.

Contents

1 Introduction

In almost all areas of life, the volume and complexity of data collected and stored are growing at a rapid pace. In the medical domain, hospitals are storing patient records, images, and test results on a regular basis. Even though this information is valuable for diagnosis and therapy, it is rarely ever used after the actual treatment and the release of the patient. In particular, it is rarely used for the extraction of information beyond individual cases, as for the improvement of medical procedures. Based on a more comprehensive analysis of such data, it may be possible to come up with new medical hypotheses as well as to find statistical evidence for existing hypotheses. This should partially overcome the limitations of traditional medical studies and trials with a small number of subjects and a fixed hypothesis to be confirmed or refuted. Overall, these new insights can help us to understand which factors may increase the risk of getting a disease, which tests provide valuable information for the status of disease, or which treatment or therapy works best for a particular patient. This knowledge can support health care workers in therapy planning or diagnosis. Powerful methods from machine learning and data mining are particularly useful for these tasks, yet need to be refined and adapted to the specific demands of medical domains.

In this thesis, we investigate techniques for test selection, subgroup discovery, and constrained clustering for the improved diagnosis in two medical domains, breast cancer and Alzheimer's disease. For the domain of breast cancer diagnosis, we mainly focus on methods for the task of test selection. This clinical task addresses the fundamental question of how to select an optimal (or near optimal) test for a patient given some background information (as demographic information or information on the patient's medical history) in order to achieve an accurate diagnosis. Or more generally, the question is how to select variables that can efficiently reduce uncertainty. This is particularly helpful to physicians who are not yet able to look back on a wealth of experience.

We introduce an approach that uses the information-theoretic concept of information maximization to guide test selection for medical diagnosis [69]. This approach yields a data-efficient method for test selection which works even on small datasets. We validate this approach on a dataset that originates from a breast cancer study of the Hospital of the University of Pennsylvania (HUP).

Since test selection may be possible just for subgroups of patients (and not all of them) as in the case of the HUP data, we investigate the use of subgroup discovery methods for this task [70]. Subgroup discovery is a data mining technique which finds the subgroups of a population that are statistically most interesting with respect to a specified property of interest. In contrast to standard applications of subgroup discovery, in the given application the property of interest is not a single variable but the relation between two variables: the outcome of a test and the actual state of disease. We tackle this by defining a so-called prediction quality function that expresses the benefit of a test (i.e., an imaging modality) for the prediction of the “correct” diagnosis (as determined by a biopsy). We propose the subgroup discovery algorithm *SD4TS* that finds subgroups of patients that share the same optimal test. Then, for a new patient, it is possible to determine the best subgroups the patient belongs to and decide for the test proposed by their majority.

The second part of the thesis deals with the domain of Alzheimer’s disease. Here, we analyze a dataset that originates from Klinikum Rechts der Isar, TU München. It consists of two different types of data: images that show the brain metabolism, and structured data, such as neuropsychologic test results or patient information. Here, the challenge lies in the combination of these diverse types of data.

We provide a method for subgroup discovery on complex outputs [46]. The method offers the possibility to correlate information from images and from structured data. In this setting, the target of subgroup discovery is not a single variable, but complex output in the form of image information. To simplify the task, we first cluster the normalized and standardized image data (PET scans), and subsequently explain the clusters in terms of psychological features using subgroup discovery. This method is applied and validated on data from patients with Alzheimer’s disease or other forms of dementia. Our experiments indicate differences in brain metabolism in terms of demographic and clinical variables.

Another possible approach to correlate image and non-image data is to apply methods for constraint-based clustering [68]. In contrast to the previous approach, in which image clusters were determined without constraints and subsequently explained by non-image data, here the image clusters are constrained by conditions on non-image data in the first place. Frequent itemsets on non-image data describe candidate clusters, which still have to be combined into an overall clustering. This combination can be controlled by a variety of user-defined constraints. The constraints may concern the type of clustering (e.g., complete clusterings, overlapping or encompassing clusters) and the composition of clusterings (e.g., certain clusters excluding others). We show that these constraints can be translated into integer linear programs, which can be solved by standard optimization packages.

Overview

This document is structured in the following way:

Chapter 2 gives a brief introduction into the two medical fields of breast cancer diagnosis and dementia research. It presents the available datasets, the clinical tasks, and a preliminary analysis of the datasets.

Chapter 3 introduces the data mining methods test selection, subgroup discovery, and constrained clustering.

Chapter 4 presents our first approach on test selection which is based on information maximization. This approach is validated on the HUP dataset on breast cancer diagnosis.

Chapter 5 shows how to use subgroup discovery for test selection with the algorithm $SD4TS$. This algorithm is validated on the HUP dataset

Chapter 6 compares the performance of the two test selection methods introduced in Chapter 4 and 5.

Chapter 7 presents how to do subgroup discovery on clustering on the dementia dataset. This approach uses the workflow that starts with clustering images and then describes the images with non-image data of the patients.

Chapter 8 introduces an alternative workflow for the correlation on image and non-image data that starts with frequent itemsets on non-image data and evaluates these sets with image data. We also provide a method for constrained clustering and show how to translate a variety of user-defined constraints into an interger-linear program.

Chapter 9 provides a summary and an outlook to possible directions for future work.

2 Introduction to Medical Datasets

The main part of our experiments is performed on two datasets that were provided to us by two large university hospitals. The first one is a data set on breast cancer diagnosis and the second one is a data set of dementia and Alzheimer's patients. In this chapter, we give a brief introduction into the two medical fields and present the stucture and some preliminary statistics of the available datasets.

2.1 Breast Cancer Diagnosis

Breast cancer is the most common type of cancerous disease among women of the western world. Approximately every tenth woman will suffer from it in her lifetime, half of whom will not survive. Even though breast cancer is such a severe disease, chances of a successful treatment and therapy are high if it is detected at an early stage. Therefore, in many countries nationwide screening[1] programs have been established. That means that every woman above a certain age is annually or biannually examined no matter if she belongs to a particular risk group or not. However, the benefit of these screening programs is controversial. Institutions like the American Cancer Society or the Deutsche Krebsforschungszentrum recommend the screening because it enables an earlier detection of cancer and therefore a more successful treatment [4, 55]. Other experts [41, 50], in contrast, investigate the disadvantages of the screening and contest its benefit: *For every 2000 women invited for screening throughout 10 years, one will have her life prolonged. In addition, 10 healthy women, who would not have been diagnosed if there had not been screening, will be diagnosed as breast cancer patients and will be treated unnecessarily. Furthermore, more than 200 women will experience important psychological distress for many months because of false positive findings. It is thus not clear whether screening does more good than harm.*[41]

Both for women that take part in a screening program and for women that require the further investigation of a suspicious finding, an accurate diagnosis is crucial in order to save lives. Besides the physical examination of the breast, there is a variety of image modalities that can help to detect cancer. In Section 2.1.1 we give a short introduction to imaging modalities for breast cancer diagnosis. The Hospital of the University of Pennsylvania designed a study specifically to investigate the performance of these modalities. In Section 2.1.2 we describe the data collected in this study and show a preliminary analysis of the data collected in this study. We introduce the clinical tasks in Section 2.1.3 and present some brief ideas for solutions in Section 2.1.4. We then address the clinical tasks of test selection in more detail in Chapters 4 - 5.

[1] Screening is an examination of patients with no signs or symptoms of the disease.

2.1.1 Imaging Modalities for Breast Cancer Diagnosis

A number of imaging modalities have been developed for breast cancer diagnosis. They include Film Mammography (FMAM), Digital Mammography (DMAM), Ultrasound (USND), and Magnetic Resonance Imaging (MRI), among others. Each has its own general weaknesses and strengths:

Film Mammography

Film mammography (FMAM) is the imaging modality most commonly used for breast cancer screening and diagnosis. FMAM is an X-ray examination of the breast. X-rays are sent through the breast causing a film on the other side to be blackened based on the densities of the tissue. This technique enables the visualization of the tissue contrasts within the breast. In general, two different projections of each breast are performed (see Figure 2.1). The craniocaudal (CC) view is taken from above a horizontally-compressed breast and the mediolateral-oblique (MLO) is taken from the side and at an angle of a diagonally-compressed breast. Due to its very high resolution, FMAM is able to detect micro-calcifications which are tiny deposits of calcium that can be an indicator for the presence of cancer.

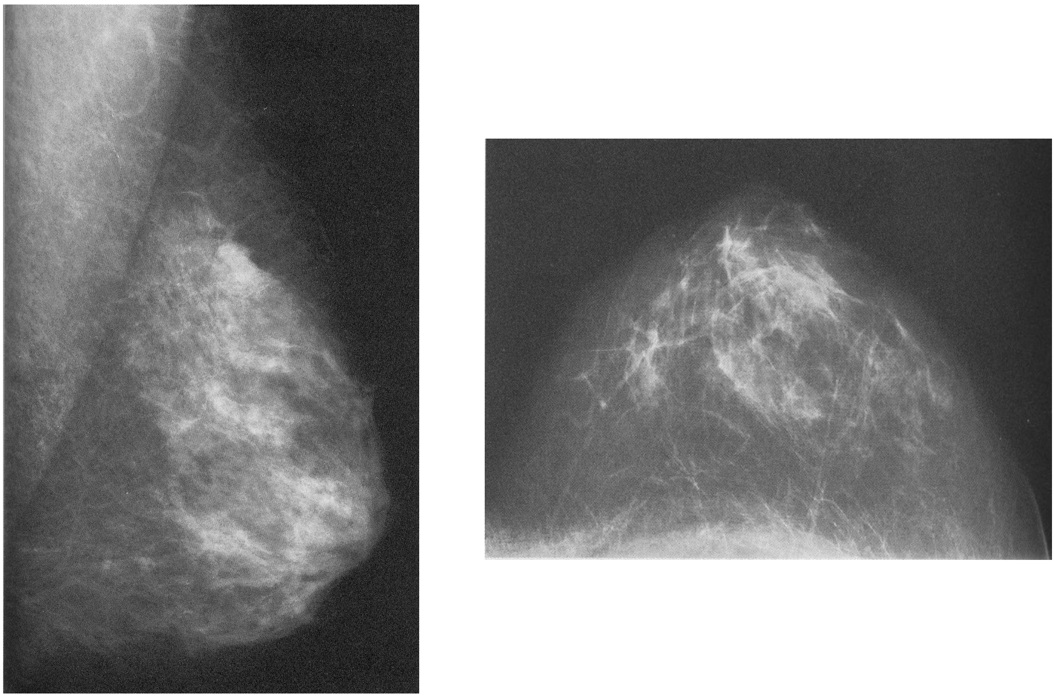


Figure 2.1: Mammogram in MLO projection (from the side) and CC projection (from top to bottom) [47].

Digital Mammography

In recent years, digital mammography (DMAM) has become more popular. Just like FMAM this technique is based on X-rays that are sent through the breast. The difference here is that the recording is done digitally and not with a film. This has the advantage of an immediate access of the image at the monitor. Also, it enables the digital postprocessing of the images (e.g, correcting the brightness or optimizing the contrast). Furthermore, DMAM provides logistic benefits, i.e., the digital image is easy to copy and distribute between different health care workers. However, the DMAM-devices are still very expensive and far less spread than FMAM-devices.

Magnetic Resonance Imaging

An radiation-free alternative to the mammographical examinations is Magnetic Resonance Imaging (MRI). Sometimes, the term Magnetic Resonance Tomography (MRT) is used because it produces two-dimensional slice images that can be composed to a three-dimensional image. For this procedure the patient gets exposed to a strong magnetic field which aligns the hydrogen atoms in the body. Then, radio waves are sent upon the protons. After switching the radio waves off, the protons reflect the obtained energy in the form of weaker radio waves. These signals are measured and then computationally transformed into an image (see Figure 2.2).

This modality especially pictures soft tissues which are high in protons. Bones, on the other hand, consist of less protons and are not visible in an MRI. In general, it is best suited for non-calcified tissue. Regions with different hydrogen-concentrations appear in different levels of gray. A discrimination between benign and malignant tissue is often possible due to the different concentration of protons. MRI provides information about the location and enhancement of a lesion. In breast cancer diagnosis, often a contrast enhanced MRI is performed. A contrast medium is injected after a native MRI scan has been done. Further scans are taken in 1-2 minutes intervals after the injection. Then it is possible to observe how the contrast medium spreads out in the tissue. To assess the difference between two scans, the radiologists make a subtraction of two scans. (The MRI scan in Figure 2.2 is an substraction). This yields a dynamic image of the breast, making it possible to judge the size and enhancement of a lesion and if it is malignant or benign.

Ultrasound

The sonography (USND) is an ultrasound-based imaging technique that can picture regions of different tissue densities in the body. For this examination, high-frequency sound waves are sent into the body. These waves are reflected or absorbed by different tissue types. The reflected sound waves are then measured and used to compute an image. The different levels of gray represent different levels of densities of the tissues (see Figure 2.2). In breast cancer diagnosis, USND is (after FMAM) the second most important modality. USND is especially useful for the detection and diagnosis of cysts

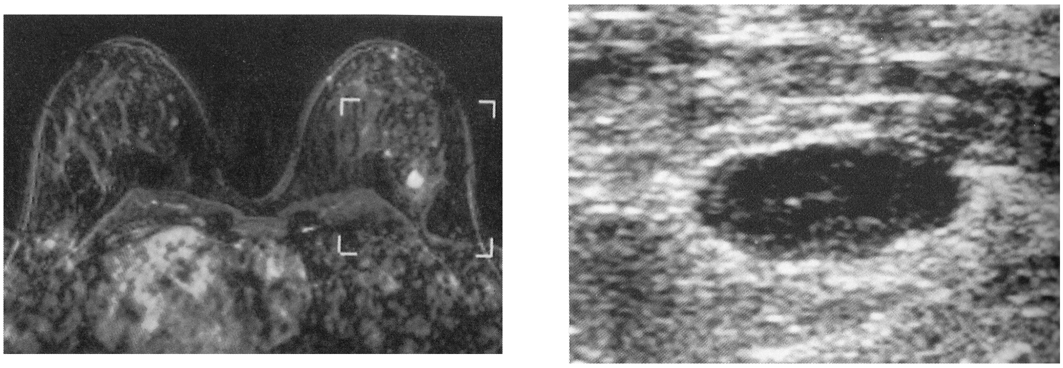


Figure 2.2: Examples for an MRI scan (left) and an USND image (right) [47]. Both images show a lesion. The white, oval, smoothly marginated enrichment in the MRI is a Fibroadenoma, which is a benign (not cancerous) breast tumor. The dark, oval, smoothly marginated lesion in the USND is a benign phyllodes tumor.

and the exclusion of carcinomas. Furthermore, it is used for patients with palpable findings that could not be visualized in FMAM due to dense tissue.

Positron Emission Tomography

Positron emission tomography (PET) is a nuclear imaging technique. This modality is especially adapted to visualize the functional processes in the body. For example, it can illustrate the metabolism activity of different body regions. For this reason, it is used for the visualization of brain activity (see Chapter 7). Chapter 7 gives a more detailed discussion of the PET functionality.

PET is a very sensitive examination which is able to visualize metabolism. Furthermore, it is not easy to anatomically map regions in PET images. Therefore, they are not used for the diagnosis of breast cancer, where it is crucial to precisely locate each lesion. However, PET is used for the staging of patients that had breast cancer. For example, it is deployed to analyze the development of lymph nodes.

2.1.2 Available Dataset

We have introduced the standard imaging modalities for breast cancer diagnosis. However, when a specific patient is under scrutiny for breast cancer, it is not clear which of these modalities is best suited to answer the basic question whether the patient has or does not have cancer. Deciding which modality to use for a particular case usually requires considerable experience of the health care professionals [47].

In 2002, the Hospital of the University of Pennsylvania (HUP) started a study to find out which imaging modalities are best suited for each patient. The study consists of three main projects:

- **Project 1** investigates the first *screening* of patients. The participants were chosen when they belong to a risk group for breast cancer (e.g., some relatives had breast cancer).
- **Project 2** focuses on the process of *diagnosis*. A patient may participate in project 2, if she had suspicious findings before.
- **Project 3** considers the *staging* of a patient, who was assessed to have breast cancer and has been treated recently.

	Project 1 Screening	Project 2 Diagnosis	Project3 Staging	Project 1&2
Patients	285	149	225	424
Lesions	715	360	629	1042
Lesions with Biopsy	139	94	211	225
Malignant	97	40	148	134
Atypia	3	6	7	9
Benign	39	48	56	82

Table 2.1: Number of available data of the different projects

The study continues until 1,000 patients have participated. Some patients take part in more than one project. We had access to the study data collected between 2002 and 2005. Table 2.1 gives an overview of the number of these cases. Since a patient can have multiple lesions, we distinguish between the number of patients and the number of lesions.

Work flow of the study

The participants underwent several examinations using five modalities: FMAM, DMAM, MRI, USND, and PET. Since PET serves only for staging project, and not for the other two projects, it is not considered in the following. Figure 2.3 illustrates the general structure of the study.

After an imaging modality examination for a patient is completed, the resulting images are assessed by an expert. The expert fills out a lesion form for each detected lesion. In the form, the lesion is characterized by several attributes like shape, size, and likelihood of malignancy. This first assessment is called *blinded*. In the data set, this information is stored in the attribute `rectype`. The blinded assessment is denoted by a `'B'`.

After the blinded assessment, the experts meet and present their assessments to each other. After this information exchange, each expert fills out a new lesion form for each identified lesion. This step is called *unblinded* and denoted by an `'U'`.

Following this assessment, they evaluate the results of the single modalities and fill out a lesion form for each lesion identified by the modality. This form is called *consensus* and denoted by a `'C'`.

Finally, they collaborate to fill out one consensus lesion form (*CLF*) per lesion summing up the information provided by the single examinations. In addition to the description of the lesion, the consensus decides if the patient should be reassessed in a follow-up examination (usually six month later) or sent to a biopsy. They also provide a recommendation on the specific modality which should be used as the guiding modality for the biopsy.

A biopsy has to be performed to get evidence of the initial assessments. That means removing a tissue sample for microscope analysis. A pathologic examination of the tissue determines whether the lesion is benign or malignant. Since biopsy is costly in terms of risks for the patient and in terms of money, it should only be performed if necessary. That is why usually only the suspicious or controversial lesions are subject to biopsy. As this study analyzes the performance of diagnosis, it is necessary to check other lesions as well, including those that are unanimously assessed as benign. Therefore, at least the index lesion of a patient is subject to biopsy. The index lesion is the lesion that was found to be suspicious in a pre-screening and brought the patient into the study. Some patients get another biopsy for further lesions.

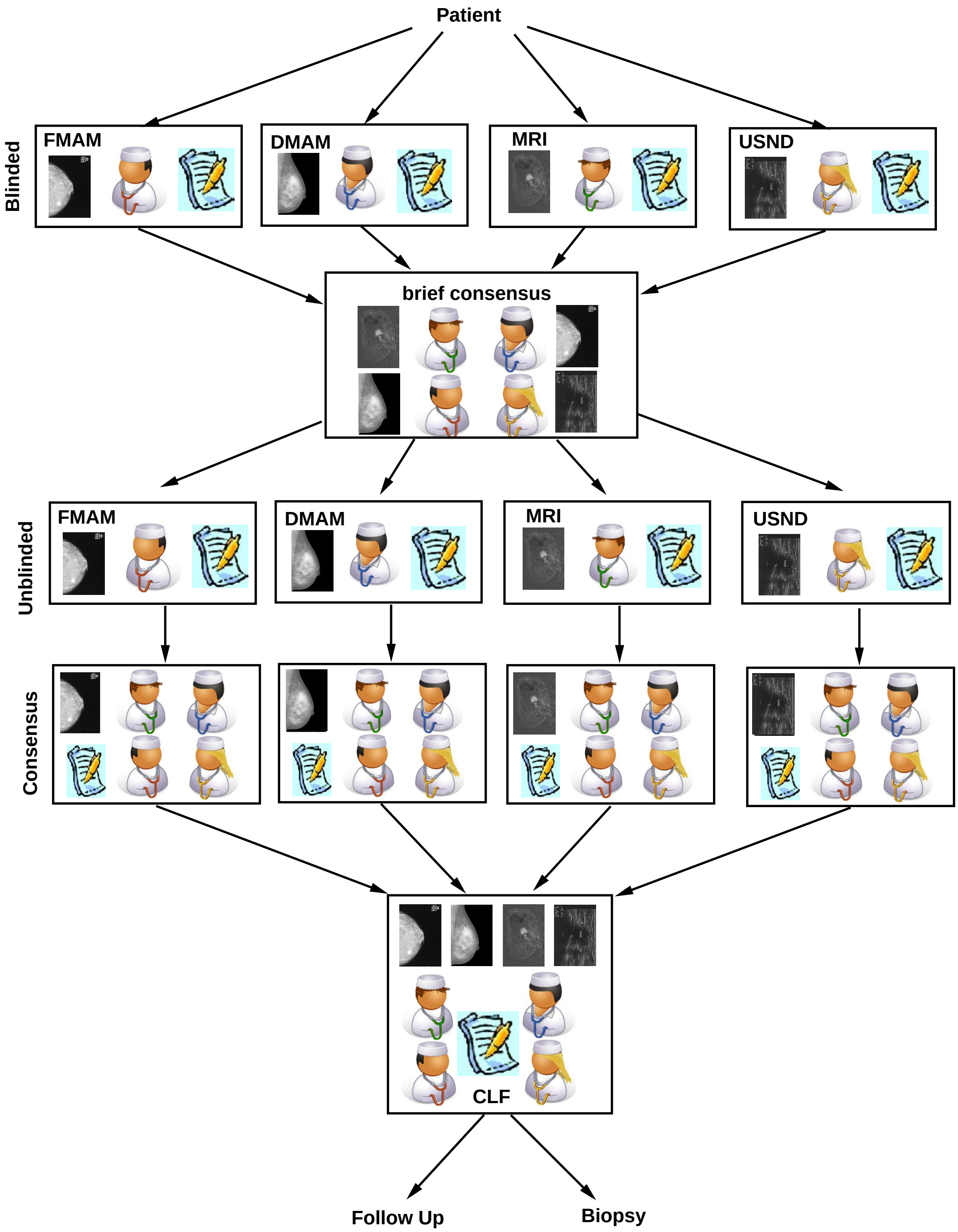


Figure 2.3: Work flow of the study

Structure of the data

The structure of the collected data is presented in the Entity Relationship diagram in Figure 2.4. We distinguish between data about the patient (Demography, Medical History, Breast Cancer Risk Summary), data that describes the lesions, and data about the examination plus their assessments.

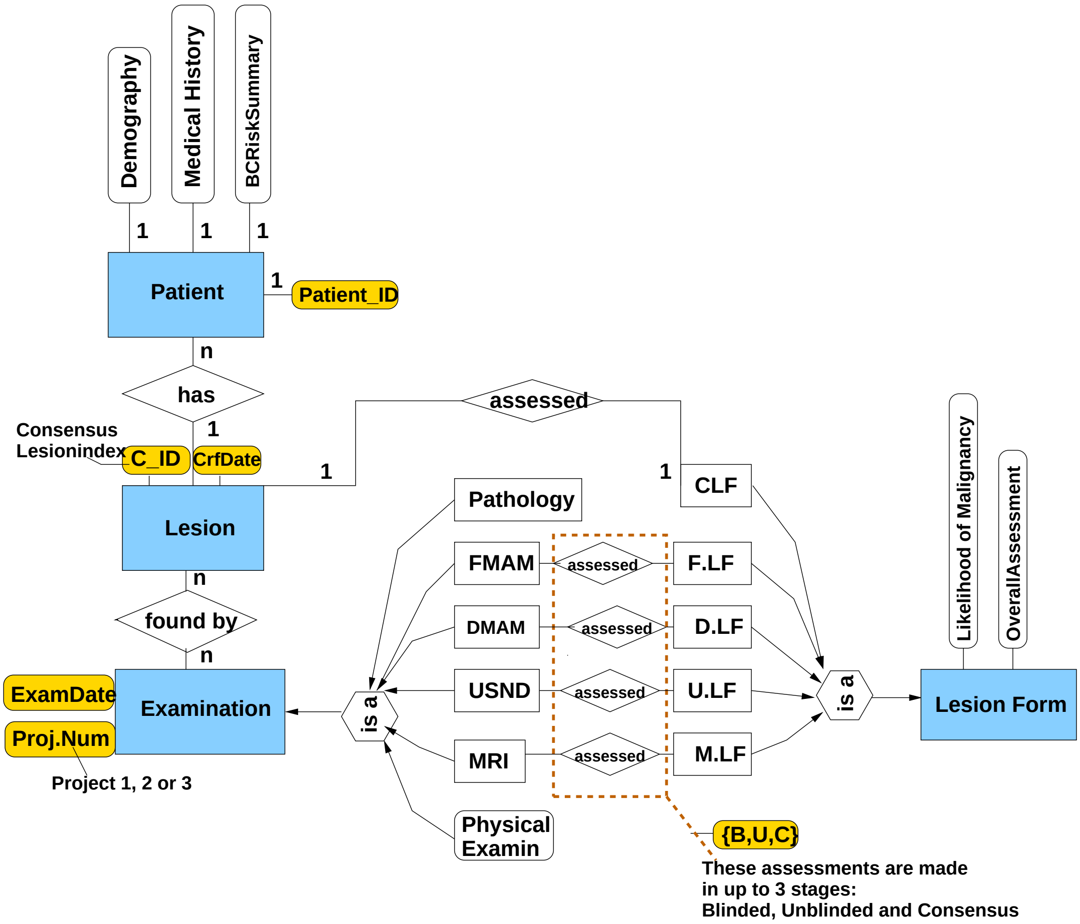


Figure 2.4: Entity Relationship Diagram of the data

Lesion indexing Each physician indexes the lesions found in the blinded assessment step. This index is called *Blinded Lesion Index*. The consensus indexes the lesions by a *Consensus Lesion Number* common to all modalities. Therefore, one is able to refer to a certain lesion unambiguously.

Multiplicity of the data The data tables covering the examinations contain a row for each `(patient, lesion, crfdate, projection number, rectype)` tuple:

- The projection number is necessary because there are up to three images per modality, each differing in the `Qualbrst` attribute, which takes one of the values {`Left, Right, Both`}.
- The `rectype` attribute takes one of the values {`B, U, C`} and indicates in which step (Blinded, Unblinded or Consensus) the assessment was made.
- `crfdate` is necessary, since a patient sometimes has a second or third examination a few months later.

Most important attributes

The diagnosis is described with two attributes in the lesion form. One is *Likelihood of Malignancy* (LOM) and the other is *Overall Assessment* (OA). LOM gives a value between 0 and 100. Higher values indicate a greater possibility of the lesion being malignant. Since the physicians tend to be conservative, they consider a lesion likely to be malignant when it has a LOM > 2. Usually, a biopsy is performed for all lesions greater than this LOM.

In addition to the LOM, each lesion gets an OA-value, which is known as the BI-RADS score (Breast Imaging Reporting and Data System) [5]. It ranges from 0 to 6 with each value having the following meaning:[2]

- **0: Incomplete**: Needs additional imaging evaluation and/or prior mammograms for comparison.
- **1: Negative**: There is nothing to comment on.
- **2: Benign Findings**: Description of findings, but no mammographic evidence of malignancy.
- **3: Probably benign**: Initial short-interval follow-up suggested. Should have a less than 2% risk of malignancy. It is not suspected to change over the follow-up interval but the radiologist would prefer to establish its stability.
- **4: Suspicious Abnormality**: Biopsy should be considered. Findings that do not have the typical appearance of malignancy, but have a wide range of probability of malignancy that is greater than those in Category 3.
- **5: Highly Suggestive of Malignancy**: Appropriate action should be taken (almost certainly malignant). These lesions have a high probability ($\geq 95\%$) of being cancerous. This category contains lesions for which one-stage surgical treatment could be considered without preliminary biopsy.

[2] Note: The HUP study considers only values from 0 to 5, because the BI-RADS score of 6 can only be given after a biopsy.

- **6:** Known Biopsy - **Proven malignancy**: appropriate action should be taken.

Preliminary analysis

Before discussing the clinical tasks in Section 2.1.3, we want to obtain an overview of the data (compare Table 2.1) and answer some preliminary questions:

1. How accurate is the overall assessment of the CLF compared to the biopsy findings? Figure 2.5 shows a histogram of the cases with the different BI-RADS scores along the x-axis. The colors indicate the biopsy result. What we expect is that all cases the CLF decides to be malignant, i.e., that have a BI-RADS of 5, are proven to be malignant in a biopsy. Vice versa, cases that are said to be benign, i.e, BI-RADS <4, are benign. The diagram shows that the majority of lesions were judged correctly. However, some of the lesions that were assessed with a BI-RADS <4 are actually malignant. This indicates that the CLF does not provide a perfect assessment.

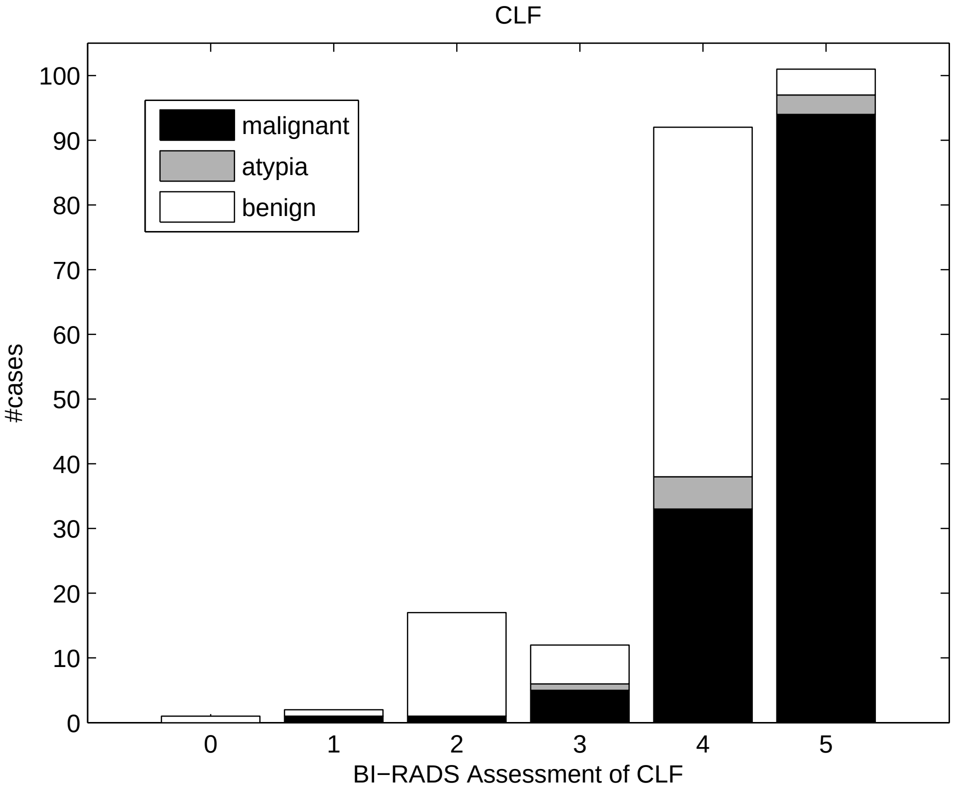


Figure 2.5: Biopsy result versus BI-RADS assessment of CLF

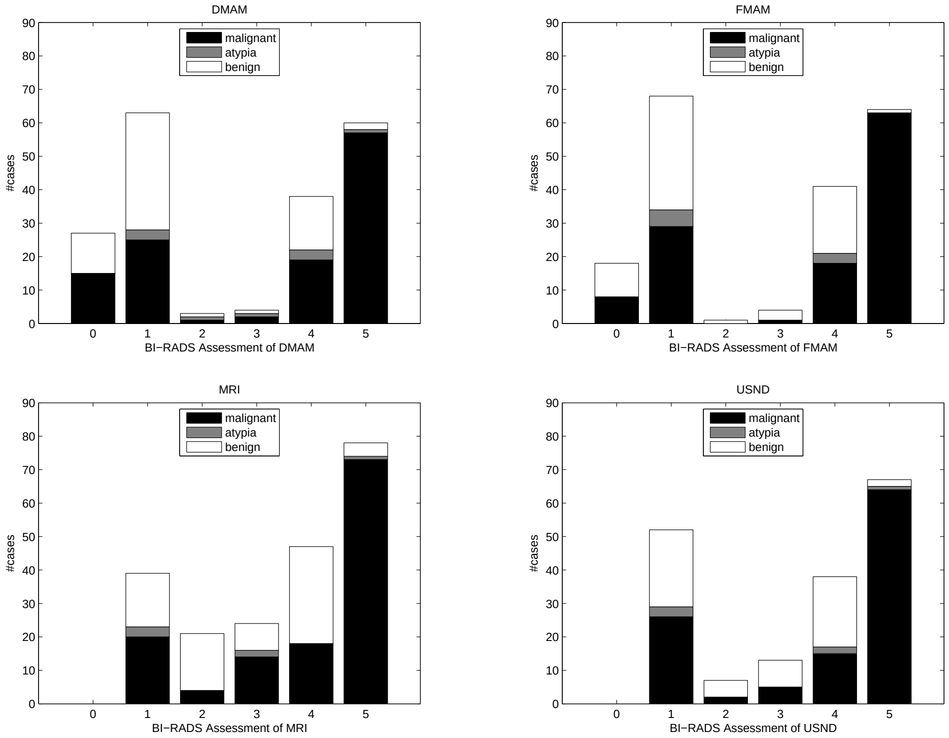


Table 2.2: Accuracy of BI-RADS assessments (x-axis) of single modalities (DMAM, FMAM, MRI and USND). The y-axis gives the number of cases of a certain x-value. The shading of the columns represents the share of lesions with the same biopsy result.

2. How accurate is the overall assessment of a modality compared to the biopsy findings?
   We now reconsider the above question for the single modality assessments. Again, we compare the BI-RADS against the actual biopsy outcome. Table 2.2 shows the resulting diagrams. It stands out that the distribution differs among the four modalities. Only FMAM and DMAM give a BI-RADS score of 0 (meaning incomplete). Also, FMAM and DMAM do not assess many lesions with a BI-RADS of 2 or 3. In MRI and USND these scores appear more often, even though the majority of those lesions turn out to be malignant. However, it is not easy to decide which modality outperforms the others. Moreover, it seems that there is no modality that does perfect for every lesion.

3. What is the difference between the four steps of assessments (Blinded, Unblinded, Consensus to CLF)? Do the Overall Assessments of the different modalities change from Blinded through Unblinded and Consensus to CLF?
   Each column of Table 2.3 presents one of these assessment steps and the rows

represent the different modalities. It is easy to see that the Blinded and Consensus assessment result in very similar plots. Therefore, we do not consider the Consensus assessment for the following analysis.

4. Are there lesions that are assessed differently depending on the modality?

Data considered in our computations

In this thesis we focus on the problem of optimal test selection for diagnosis. Therefore we work with the data collected from the Screening and Diagnosis project. We do not use the data from the Staging project because this data can refer to patients that had their primary cancer removed after film mammography but before the other modalities. Furthermore, since we are interested in the information originating from single imaging modalities, we consider only the blinded assessments and do not take into account unblinded and consensus assessments. Our preliminary analysis showed that Likelihood of Malignancy is a very imprecise measure. Therefore, we consider the Overall Assessment attribute to contain the actual diagnosis. For all our experiments we consider only lesions with biopsy results, which we use as the true disease status.

Missing values For some lesions we have no entries in the submitted .csv-file (CRF Lesion Form). It is not clear if the lesion was not detected by the examination or if the examination was not completed for the patient. If a lesion l of a patient p is recorded in the lesion form of modality A, and not recorded in the lesion form of modality B, we distinguish two different cases:

1. There is a lesion l_2 of patient p recorded in the lesion form of B, or
2. modality B has no recordings of lesions of patient p.

In the first case, we know that the examination B was done for patient p. Therefore, we can conclude that the lesion l was not detected by the modality, and can assume BI-RADS=1 (no findings). (Of course, it could also be the case that the data of lesion l was just not entered into the database or was entered, but not submitted to us yet.) In the second case, we have no evidence that the examination B was done for patient p, and we therefore assume that the examination was not done or that we do not yet know its result. In this case, we consider the BI-RADS score of B for lesion l as missing.

2.1.3 Clinical Tasks

The physicians are interested in answers to the following:

1. Determine for a given lesion or patient which modality performs best. This task is needed for two scenarios. In the first scenario we have only knowledge about the patient and her medical history. In the second scenario we additionally have knowledge about the outcome of a previously performed test. In the latter case, the task is to determine which modality to choose next, or rather if we have to consider further modalities at all.

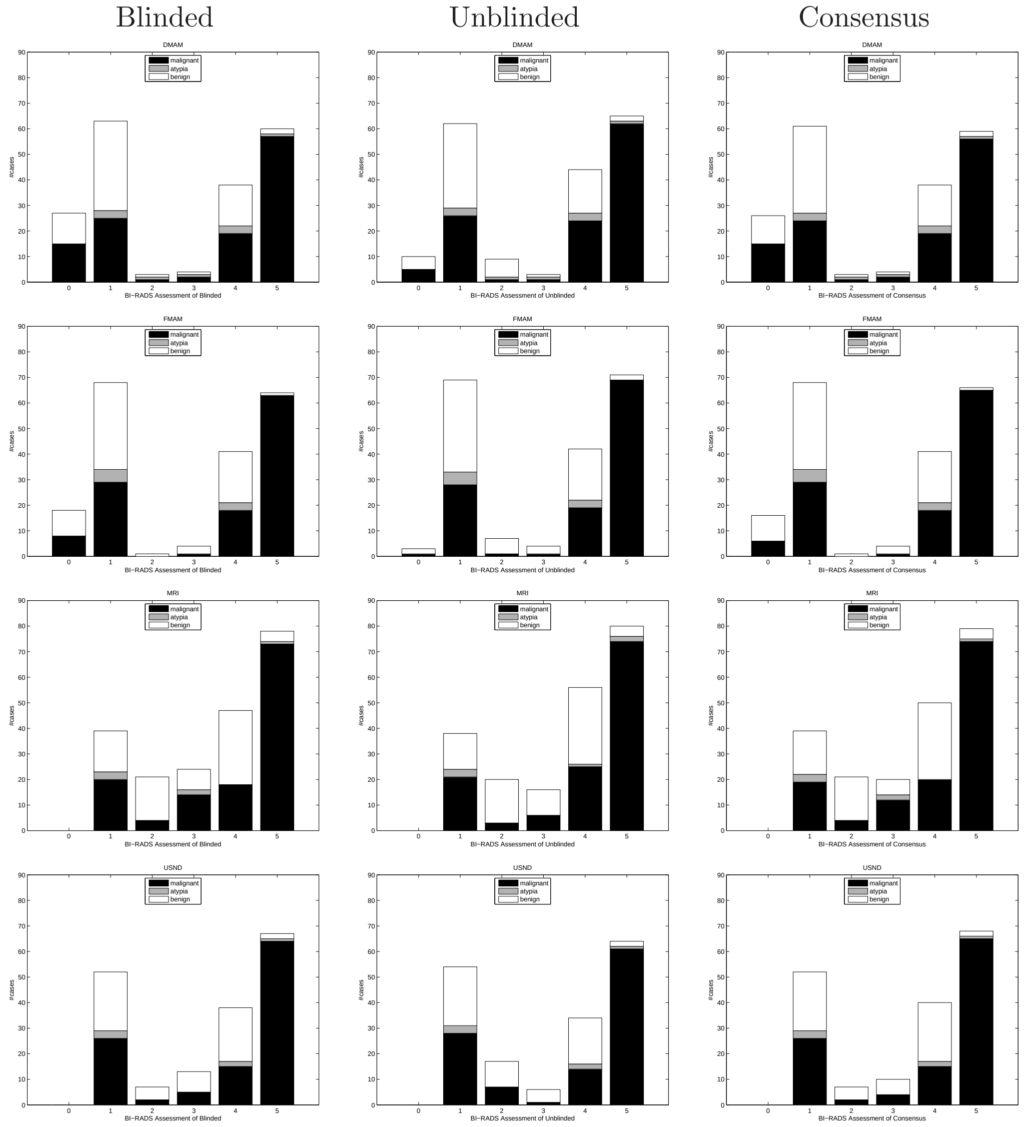


Table 2.3: Difference between the three stages Blinded, Unblinded, Consensus Assessment of the four modalities: Accuracy of BI-RADS assessments (x-axis) of single modalities (DMAM, FMAM, MRI and USND). The y-axis gives the number of cases of a certain x-value. The color of the columns represents the share of lesions with the same biopsy result.

2. For each situation determine which of the following is best:
   - do nothing (lesion is benign),
   - do a biopsy (lesion is malignant), or
   - do another test (MRI, USND, DMAM or FMAM).
3. Is the concept of a CLF adapted for the diagnosing process? Are there better ways to combine data from different tests? Prof. Mitch Schnall has an interest in finding better ways to combine data from different modalities. Today, the standard approach is based on a simple consensus of physicians. However, it is not clear whether this procedure is most effective for diagnosis.
4. Given a new case, find similar (useful) cases treated in the past. Knowing their diagnosis and staging data could help diagnosing new cases more accurately.

In this thesis, we set the focus on solving the first task. In chapter 4 and 5 we introduce two methods how to approach the task of optimal test selection, present the results, and discuss their performance. Ideas how to solve the remaining clinical tasks are presented in the following section.

2.1.4 Brief Ideas for Approaching the Clinical Tasks

We briefly sketch some ideas how to solve the clinical tasks 2-4 based on the given data set. The implementation of these ideas is beyond the scope of this thesis.

Decision algorithm

The approaches to test selection presented in chapter 4 and 5 can be extended to a decision algorithm that can determine, for a given situation, which of the following actions has to be taken:

- do nothing (benign),
- do a biopsy (malignant), or
- do another test (MRI, USND, DMAM or FMAM).

The basic idea is to start with an initial set of lesions S. For each lesion $s \in S$ we attempt to learn the best diagnostic pathway, i.e., to determine the modalities that should be applied in order to arrive at an accurate diagnosis. After the training, we can determine an optimal diagnostic pathway for the new lesion $s' \notin S$.

Since there is no globally optimal modality for the complete set of examples, we consider different diagnostic pathways for different lesions. We need a partition of the lesion set in a way that all lesions in a subgroup follow the same diagnostic pathway. The simplest way to determine a partition is to select an attribute Z and divide the given set S into n subgroups $S_{z_1}, ...S_{z_n}$, where S_{z_i} is the subset of S for which the picked attribute Z has the value z_i. Let X_j be a modality with $j \in \{$DMAM, FMAM, MRI, USND$\}$, $S' \subseteq S$ and $G(X_j|S')$ be an arbitrary function, that expresses the benefit for the accurate diagnosis of the lesion set S' provided by applying modality X_j. Examples for function G are mutual information (see Section 3.1.1) or prediction quality (see Section 5.2.1). An attribute Z is suitable to split the set S if one of the following holds:

a) The choice of the best modality depends on Z:
$\operatorname*{argmax}_j G(X_j|S_{z_1}) \neq \operatorname*{argmax}_j G(X_j|S_{z_2})$

b) The partition leads to a better modality performance:

   - or in general: $\frac{1}{|Z|} \cdot \sum_{z_i \in Z} \max_j G(X_j|S_{z_i}) > \max_j G(X_j|S)$
   - or for a particular attribute value:
     $\frac{1}{|Z|} \cdot \max_{z_i \in Z} \max_j G(X_j|S_{z_i}) > \max_j G(X_j|S)$

c) We can optimize the maximum (or average) of G over the four modalities to determine the best splitting attribute Z^*.

   - $Z^* = \operatorname*{argmax}_Z \sum_{z_i \in Z} \max_j G(X_j|S_{z_i})$

- $Z^* = \underset{Z}{\operatorname{argmax}} \max_{z_i \in Z} \max_j G(X_j|S_{z_i})$

The stopping criterion of the algorithm is reached when either one of the modalities performs well enough, i.e., $\max_j G(X_j|S) > c$ for a user-defined c, or one modality performs significantly better than the others. The algorithm will return this modality.

Combination of information from multiple sources - learning the CLF

The CLF (Consensus Lesion Form) is the final conclusion of the four physicians after having discussed the outcome of all four examinations. For each lesion, they describe its position (e.g., radial location, distance from nipple at skin, depth from skin at tangent). Then, they agree upon a LOM and an OA assessment. Furthermore, they can give a recommendation to do a follow-up examination or a biopsy. In the latter case, they name the best guiding modality.

The preliminary results (Section 2.1.2) showed that the CLF OA differs from the biopsy findings. Therefore, we cannot consider it as ground truth. Since every lesion has a CLF assessment whereas only few have a proven biopsy, there is a big interest in learning the rationale of the CLF assessment. Especially, it is helpful to learn to predict the CLF assessment by combining the assessments of the different modalities. In a second step we could improve its accuracy regarding the biopsy findings. Then, we obtain a ground truth label for all lesions with an CLF assessment. Thus, we can use a larger number of data points for the clinical tasks investigated above.

One approach for this task is to learn the final diagnosis (CLF OA) given the OA and LOM of all other exams and some background attributes about the patient and the lesion. Knowing how the CLF OA is determined can help finding better assessments. At least, it will be possible to show if a certain modality has a stronger influence on the choice of CLF OA than it should have.

Further steps will include an analysis of the performance of CLF OA compared to the biopsy and of the outcome of the learned classifier compared to the biopsy. The analysis results can be used to restrict the training set for the classifier to the examples that have a good CLF OA value regarding the biopsy. This classifier will then be applied to predict the examples with an worst CLF OA. This could lead to a general improvement of the CLF OA accuracy.

A further approach could be to skip the CLF and combine the information obtained in the single modalities directly to infer an assessment, which should be close to the findings of the biopsy.

Another option is to learn which guiding modality is selected given breast density, calcification and the four Overall Assessments (plus the CLF data). There is no proof for the performance of that choice. The task of interest would be to determine the actual decision criteria of physicians.

Furthermore, an aim is to investigate the location of a lesion. It is interesting to see if the description of the location varies between the different modalities. It might be possible to determine reasons for variations.

Case-based reasoning

Physicians often relate new cases to be diagnosed to cases seen in the past. Case-based reasoning approaches and systems attempt to mimic this form of reasoning, partially on the basis of similarity or dissimilarity (distance) measures. Baumeister *et al.* present an inductive approach to learn similarities for case-based reasoning in the field of medical diagnosis [14]. For a more general intoduction to case-based reasoning refer to Aamodt *et al.* [1] or Watson *et al.* [89].

The definition of a suitable similarity measure in the context of the HUP study raises a couple of questions: Are two lesions similar if they have the same modality performance (the same modality performs best for both), or are they similar if they have the same lesion- or patient-based features (e.g, same size, shape, location, or same breast density, medical history, demographic features)? The same biopsy finding could define similarity as well.

Distance matrices as similarity function The difficulty is to develop a similarity function without preprocessing the data. If we have a list of given cases $y_1, y_2, \ldots, y_n$ and a new case x arrives, we want to select the most similar y_is of the given cases. Let us assume we have a distance $m \times m$-matrix W and the vector $y_i = (a_1, \ldots, a_m)$ with m the number of attributes of a case. Then, we obtain the most similar case as $arg \min_i xWy_i$.

W should be chosen in order to maximize the similarity between cases with the same biopsy outcome and minimize the similarity between cases with different biopsy results.

$$W^* = arg \min_W \sum_{i,j Bio_i = Bio_j} y_i W y_j - \sum_{i,j Bio_i \neq Bio_j} y_i W y_j$$

We can simplify the equation by defining a matrix $B = \begin{cases} 1 & \texttt{if } y_i = y_j \\ -1 & \texttt{otherwise} \end{cases}$ and obtain:

$$W^* = arg \min_W \sum_{i,j} y_i \cdot B \cdot W \cdot B \cdot y_j$$

To find the minimum, we first need to define the derivative and than find its null.

Use of Unlabeled Data

The biopsy outcome provides the ground truth of the lesions malignancy (labels). Most clinical tasks refer to this ground truth. However, the majority (about two thirds) of the lesions does not get evidence through biopsy. It is an open question how this unlabeled data (i.e., data without biopsy) can be incorporated into our approaches. In order to solve this, we need to discuss the following questions. Is the selection of which lesion gets a biopsy a random selection? Is there a reason why the unlabeled lesions do not have a biopsy? For instance a possible reason for not getting a biopsy is that the physicians are sure about their results, or an adjacent lesion of the patient already gets a biopsy.

Surprisingly, 26% of the lesions without biopsy got a recommendation in the CLF to do a biopsy, and only 60% of the lesions that got a recommendation for doing a biopsy, de facto got a biopsy.

Depending on the answers to the previous questions, different approaches could be taken: One where a random distribution of labeled and unlabeled data is assumed, and another where the index lesion is assumed to be the one with the highest probability of being malignant, and the extra lesions with biopsy are selected because they are likely to be malignant as well.

If the biopsy result is not used as a target variable, standard approaches for missing values from machine learning and data mining could be applied. If the biopsy result is used as a target variable, however, methods could benefit from the unlabeled data as well. For instance, statistical approaches could use the unlabeled data to obtain a more realistic joint probability distribution of the variables of interest. Alternatively, one could apply methods from semi-supervised learning [22, 95, 96].

In this context, open questions include the following: Can you include information of neighboring lesions, if multiple lesions exist in order to determine their label? Can we assume that the overall assessment of the CLF is accurate enough in order to obtain the actual status of disease for the unlabeled data?

2.2 Alzheimer's Disease

Every year 4.6 million people are diagnosed worldwide with some type of dementia, 200,000 alone in Germany [3, 12]. Due to the "aging society", this figure is expected to increase continuously over the next decades. So experts estimate the number of 30 million dementia patients in 2008 to raise to a number of 100 million patients in 2050. Even though 98% dementia patients are 65 years and older, younger people are indirectly affected when they have relatives with dementia. Above an age of 85, almost every fourth person suffers from dementia (compare Table 2.4).

Age	Annual incidence per 100		Prevalence (%)	
Group	Males	Females	Males	Females
60-64	0.2	0.2	0.4	0.4
65-69	0.2	0.3	1.6	1.0
70-74	0.6	0.5	2.9	3.1
75-79	1.4	1.8	5.6	6.0
80-84	2.8	3.4	11.0	12.6
85-89	3.9	5.4	12.8	20.2
90+	4.0	8.2	22.1	30.8

Table 2.4: Incidence and prevalence rates of dementia from the EURODEM meta-analysis carried out in the 1990s in eight different European countries [60]. Prevalence is an estimate of the proportion of individuals in a population that have a disease at one point in time. The annual incidence rate gives an estimate of the number of individuals that are diagnosed with a disease within a year.

Dementia is the name for diseases that are based on the loss of mental abilities, such as memory, learning, reasoning, language, and orientation. Eventually, dementia makes it impossible for the patient to manage every day life independently . There are many causes for dementia. Alzheimer's disease (or Morbus Alzheimer) is the most common one being responsible for 60% of the dementia cases. Other forms of dementia include dementia with Lewy bodies, vascular dementia, frontotemporal dementia, or dementia as a secondary disease of diseases like Parkinson's or Pick's disease. Many patients are affected by a combination of different types of dementia. Although some forms of dementia like Alzheimer's disease are well-known and characterized for one hundred years now, the underlying mechanisms are still not sufficiently understood. Therefore, there is great interest in expanding our knowledge of various forms of dementia, including Alzheimer's disease.

The first case of Alzheimer's disease (AD) was described by the German neurologist Alois Alzheimer in 1906. By this time, persons showing symptoms of dementia were considered to be insane. Alzheimer was the first to examine the brain of those patients post-mortem. This let him discover, that the psychological changes of the patient were

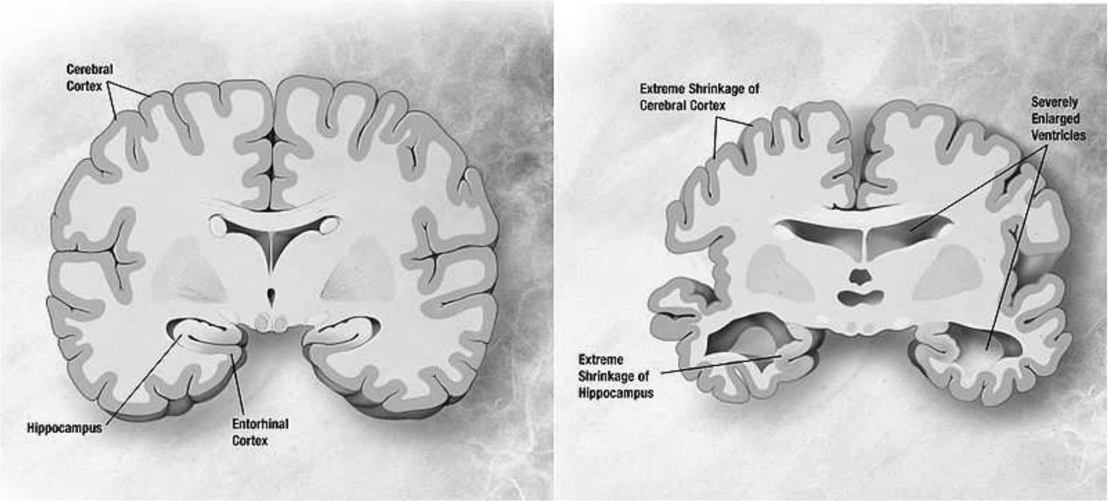

Cerebral Cortex
Hippocampus
Entorhinal Cortex
Extreme Shrinkage of Cerebral Cortex
Severely Enlarged Ventricles
Extreme Shrinkage of Hippocampus

Figure 2.6: Healthy brain versus brain with Alzheimer's disease [71].

caused by an actual degeneration of the brain. From this time, dementia was accepted as an actual disease.

Figure 2.6 shows how a brain of an Alzheimer's patient looks compared to a healthy brain. The cerebral cortex as well as the hippocampus show a strong shrinkage compared to the healthy brain. Furthermore the ventricles are enlarged.

Causes

Even today, more than one hundred years after the first discovery of the disease, the causes of Alzheimer's are not completely understood. The main theory is that the symptoms of Alzheimer's disease are caused by the deposition of pathological proteins in the form of intracellular tangles and extracellular plaques. This deposition is followed by neuron death and deficits in neurotransmitter systems. Furthermore, deficits in glucose metabolism occur.

Therapy

Till now it is not possible to treat the causes of the disease and completely cure the patient. It is still a terminal disease and there is still no way to stop the process of degeneration. So most forms of therapies aim for a slow-down of the process of degeneration. Which means that the patients can manage their daily lifes for a longer time. Types of medications that were proposed for this purpose include acetylcholinesterase inhibitors, memantine and atypical neuroleptics. However, the effect of these medications is controversial. Since there are both studies that show that they do and do not significantly slow down the course of disease. Another way of therapy is of psychosocial nature. For instance, it is known that patients benefit from a structured day and a simplified environment.

2.2.1 Methods for Diagnosis

When a person shows symptoms of dementia, at first a general neurological and physical examination is performed. Furthermore, relatives of the patient are questioned to describe the patient's behavior in every day situations. This is a valuable source of information for the diagnosis since many patients cannot judge themselves appropriately anymore. Those patients that have suspicious test results, additionally, undergo radiological examinations that make the actual degeneration of the brain visible. After all examinations the physicians diagnose the patient. The diagnosis is indicated with a ICD code [90] (compare Table 2.5). The diagnosis posed by the physicians is the basis for therapy and treatment of the patients. However, a definite diagnosis of AD can only be given by a post-mortem examination of the brain. Of course, this is too late for proposing a therapy or treatment. Therefore, in recent years a lot of dementia research focused on determining biomarkers that are indicative for different types of dementia. This will help gaining a higher certainty of the diagnosis.

Disease	ICD code
Mild cognitive disorder	F06.7
Alzheimer's disease	F00.0, F00.1, F00.2
Depression	F32.X, F33.X
Other forms of dementia	F00.X - F09.X

Table 2.5: Categories of ICD codes for dementia

In the following we will introduce the most common examinations.

Neuropsychological Tests

We start with a description of three common neuropsychological tests, CERAD, CDR and clock test.

The standard examination to judge the cognitive impairment of a patient is called *CERAD* (Consortium to establish a registry for Alzheimer's disease) [67]. Since its establishment in 1986 by the National Institute on Aging, it is widely used to evaluate and diagnose patients with AD. CERAD provides a standardized way to judge the mental functionalities that can be affected by dementia: semantic memory, word finding, visual cognition, orientation, concentration, direct retentiveness, visuo-construction and delayed retentiveness The different functionalities are addressed in eight subtests and the achievable scores in parentheses:

1. Mini-Mental State Examination (MMSE) (0 - 30)

2. Verbal fluency (0 - ∞)

3. Modified Boston Naming Test (MBNT) (0 - 15)
4. Construction praxis (0 - 11)
5. Learning of word lists (0 - 30)
6. Recall of word lists (0 - 10)
7. Recognition of word lists (0 - 20)
8. Constructional praxis recall (0 - 11)

The healthier a person, the more points are expected. It is possible to calculate a score (0 - 100) over all CERAD subtests [21]. Depending of the severity of dementia, it takes between 30 and 50 minutes to complete the CERAD test battery. In some scenarios, it is sufficient to only do the MMSE since it provides already a broad survey of the cognitive impairment. For instance, the MMSE is often used for a quick follow-up examination of previously diagnosed patients.

The Clinical Dementia Rating test (*CDR*) [48, 66] is used to judge how well the patient can handle the daily life. The CDR is an interview of the patient or a relative of the patient and refers to the following subcategories:

1. Memory
2. Orientation
3. Judgement
4. Social
5. Hobbies
6. Personal

For each category the scores range from 0 (no deficits) to 3 (severe deficits). Furthermore, a CDR global score is determined by an algorithm that combines the scores of the subcategories.

The third test that is performed in the *clock drawing test*. Here, the patient is asked to draw a clock (including the clock face) that shows ten past eleven. This simple task requires complex skills as orientation, construction ability, concentration, and short term memory. Therefore, the results provide a valuable information on how severe the patient is affected by dementia. The clocks are graded in a range from one to six depending on how well the clock is drawn, and if it shows the correct time. A score of one reflects the perfect clock and a score of three and above is an indicator for dementia. Figure 2.7 shows the clocks drawn by different patients.

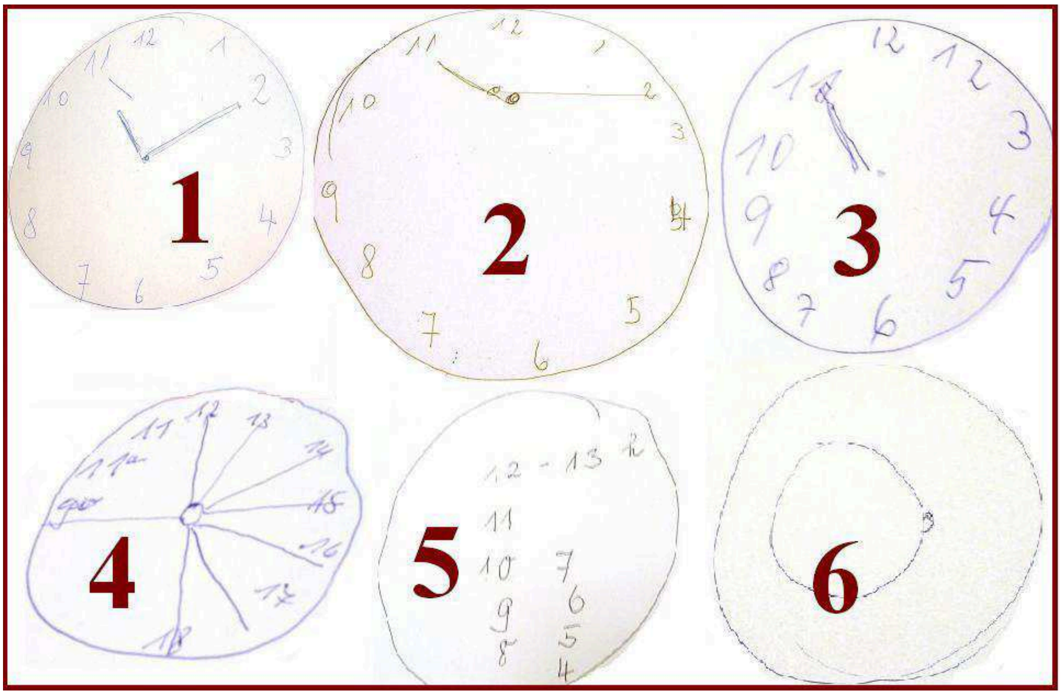


Figure 2.7: Results of the clock test with scores from 1 (very well) to 6 (poorly)

Radiologic examinations

To evaluate the degeneration of the brain it is often necessary to illustrate the brain with the help of imaging modalities. Most commonly used are MRI, CT, and PET. In this thesis we mainly focus on PET scans.

PET is a non-invasive medical imaging procedure that has been used for diagnostics of dementia since the early eighties. It displays a three-dimensional map of the glucose metabolism of the body and is based on the decay of radioactive markers, which are injected into the patient. A scanner records the cell activation, and a computer calculates the three-dimensional image of the metabolism. The recorded PETs indicate the metabolism of the brain, i.e., the transformation of glucose. This reflects the activity of neural cells. The brain of patients suffering from dementia contains regions, where the metabolism is clearly lowered. In Alzheimer's disease, the pattern of hypometabolism starts at the hippocampus and spreads over the entire cortex as the disease progresses, sparing few areas such as the motor and primary visual cortices.

2.2.2 Available Dataset

The data was provided by the psychiatry and nuclear medicine departments of Klinikum rechts der Isar of Technische Universität München. It consists of demographic information, clinical data, including neuropsychological test results, and PET scans showing the patient's cerebral glucose metabolism. We had access to clinical and demographic data of 4,037 patient visits and 454 PET scans that have been collected between 1995 and 2006.

To increase the quality of the data, we revised the existing psychological data of the 1,100 visits of patients having a corresponding PET or cerebrospinal fluid examination (test for certain protein levels). Our revision included the correction of typing errors and the completion of electronically available test results. For some patients with PET, a revision was not possible due to missing patient records. 257 PETs belong to patients with revised psychological and demographic values. The overall effort for revising the data was approximately four person months.

Demographic Data

We had access to demographic data of 4,037 patient visits (751 with revised data). The data covers gender, age, years of education, type of graduation and profession.

Neuropsychological Data

The clinical data for each patient consists of both psychological test results and information about which of the other tests (e.g., MRI, CT, SPECT and EEG) were completed during a hospital visit. Additionally, the diagnosis for each visit is provided in the form of ICD-10 codes [90].

The assessment instruments are the standard tests to diagnose dementia: Consortium to Establish a Registry for Alzheimer's Disease Neuropsychological Assessment Battery (*CERAD*), Clinical Dementia Rating (*CDR*), and the clock drawing test *CDT*. These tests are optionally done at each patient visit at the physician's discretion.

Preprocessing of PET Scans

Before physicians can actually use the PET scans, the images have to be processed by a sequence of transformation steps. The image preprocessing pipeline of our study is illustrated in the upper left-hand part of Figure 7.1. Due to measurement irregularities and the motion of patients during the recording of the images, each image has to be rotated and translated, such that they all fit into the same template. This is achieved by SPM5[3]. Subsequently, the images are forwarded to (X)MedCon[4], which transforms the data into raw ASCII files representing the intensity of each voxel as a real value. The last step is the normalization by dividing each voxel value by the mean voxel value of the image. At the end of the preprocessing, each file consists of 69 matrices of 79 rows and 96 columns, summing up to 523,296 voxels. Each of the 69 layers reflects a horizontal cut through the brain, presenting a two dimensional image of the layer with voxels displaying the intensity of metabolism of the corresponding region.

In order to visualize a group of scans, we computed a *mean image* using SPM5. In a mean image, each voxel contains the mean voxel value at that position. The mean image

[3] SPM5 release 12/01/2005, based on Matlab 7.3, see http://www.fil.ion.ucl.ac.uk/spm/software/spm5/

[4] XMedcon 0.9.9.3, http://xmedcon.sourceforge.net/

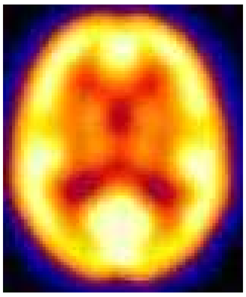


Figure 2.8: Mean image of 20 healthy controls.

has the advantage of showing the overall properties of a group. Note that each mean image shown in this thesis displays only one layer of the three-dimensional mean image. We always choose the same layer (layer 32) to facilitate a comparison of the images. However, for the evaluation of clusters, we take into account all 69 layers.

Figure 2.8 shows the mean image of 20 healthy controls. The "north" displays the frontal region of the brain from an top-down view. The lighter the spots, the more active is the underlying tissue. Both sides of the brain are symmetric in their metabolism. The lateral regions have an increased metabolism, while the center has lower metabolism. This is due to the fact that the brain cells are residing in the outer areas of the brain, whereas the brain center mostly consists of dendrites.

2.2.3 Clinical Task

One of the goals of medical research in the area of dementia is to correlate images of the brain with the clinical test results. We address this problem in the chapters 7 and 8 and introduce two approaches that provide a correlation.

3 Introduction to Data Mining Methods

In this chapter we introduce the data mining methods used in this thesis. These are test selection, subgroup discovery, and constrained clustering.

3.1 Test Selection

Test selection is the task of determining an optimal test for a given situation. In this thesis we focus on test selection in the context of medical diagnosis. During patient care, health care workers are constantly faced with the problem of arriving at a diagnosis decision or action plan for diagnosis at multiple time instances, such as after observing symptoms of a patient or after gathering new clinical findings from laboratory tests or other sources. These decisions are critical when assigning a proper treatment to a patient. However, in most cases, there is no unique and clear diagnosis or obvious action plan. Even after patient history has been gathered and some tests performed, for many cases there can still be considerable uncertainty about the correct diagnosis. More generally, at any point in the diagnosis process the clinician is faced with numerous questions or options regarding what the best course of action should be in order to properly arrive to the correct diagnosis with efficient use of resources; alternatively, to determine that sufficient information has been collected so that a reliable diagnosis can be rendered. Note that the physician faces the same challenges while making decisions about therapy (for example, whether the therapy should be picked now, or more information should be collected).

More generally, diagnosis is the art or act of identifying a disease from its signs and symptoms. This implies that the more information is available about a patient, the easier it is to pose an accurate diagnosis. Information can be obtained by a variety of tests including questioning the patient, physical examinations, imaging modalities, or laboratory tests. However, due to costs, time, and risks or discomfort for the patient, in clinical routine it is often preferable for patients to undergo as few tests as needed. Consequently, there is a trade-off between the costs (and number) of tests and the accuracy of the diagnosis. Therefore, optimal test selection plays a key role in diagnosis.

From an abstract point of view, the situation is similar to many machine learning applications in practice, where only a fraction of the data is available for processing, while (potentially useful) additional data can be gathered on a limited basis. A key question in these applications is how to best decide what data should be gathered or observed. Considering the medical domain again, data collection may involve exposing the patient to high-risk procedures or tests, and thus it seems reasonable to assess the medical *value* of such procedures beforehand for the particular patient.

Test selection has been investigated extensively in the medical and the statistical literature [7, 33, 40, 80, 93]. In the following, we will present some of the existing approaches. The two most established measures for the evaluation of tests are Shannon entropy and Gini index.

3.1.1 Shannon Entropy - Mutual Information

The first measure originates from the area information theory and was introduced in 1948 by the mathematician Claude Shannon [83]. In order to describe the concept, we first introduce our nomenclature: Let Y be the diagnostic variable modeling the status of disease. The possible values of Y are denoted $y_1, \ldots, y_n$. Let X_j be the test variables with $j \in \{1, \ldots, m\}$ and the results of test X_j denoted x_{jk} with $k \in \{1, \ldots, m_j\}$ and $m_j \geq 2$.

The *Shannon entropy* $H(Y)$ is the expected amount (in bits) of information required to determine the value of Y with certainty. The entropy is thus a measure of uncertainty. It is defined as:

$$H(Y) = -\sum_{i=1}^{n} P(y_i) \cdot \log_2 P(y_i)$$

For varying probability distributions of a binary variable of interest Y, the entropy is illustrated in Figure 3.1.

Let us assume that a test X_j has been performed yielding the result x_{jk}. This additional information will change the probability distribution over Y to the posterior distribution given $X_j = x_{jk}$. The Shannon entropy will then change to the conditional entropy:

$$H(Y|X_j = x_{jk}) = -\sum_{i=1}^{n} P(y_i|X_j = x_{jk}) \cdot \log_2 P(y_i|X_j = x_{jk})$$

Since we do not know the outcome of test X_j prior to performing it, each possible result x_{jk} is yielded with a probability $P(x_{jk})$. Therefore, the expected entropy of the posterior probability is:

$$H(Y|X_j) = \sum_{k=1}^{m_j} P(x_{jk}) \cdot H(Y|X_j = x_{jk})$$

If we consider now a particular test X_j, we want to know to what degree the test result is expected to decrease the uncertainty about Y. This value is expressed by the *mutual information* (MI) of X_j and Y: $I(X_j, Y)$. Mutual information is also called *information gain*, since it is nothing else than the reduction of the entropy H of the distribution over Y.

$$\begin{aligned} I(X_j, Y) = &\; H(Y) - \sum_{k=1}^{m_j} P(x_{jk}) \cdot H(Y|X_j = x_{jk}) \\ = &\; H(Y) - H(Y|X_j) \end{aligned}$$

In other words the MI describes the amount of information that a test variable X_j provides about the variable Y. It determines how well X_j splits the set of examples S in order to determine the outcome of a target attribute Y like the status of disease. MI

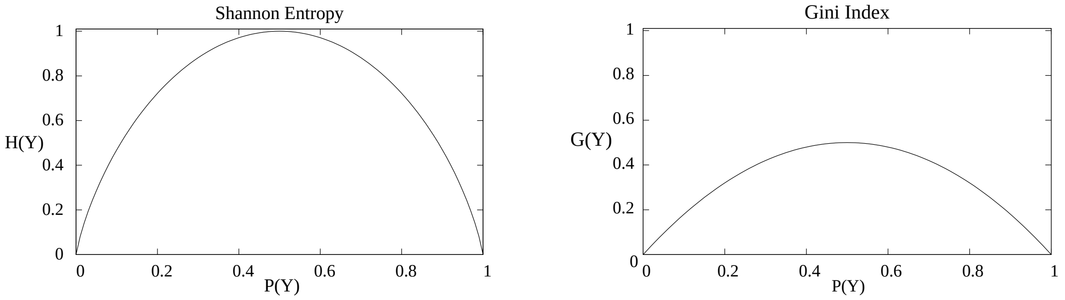


Figure 3.1: Shannon entropy and Gini index for a binary variable of interest Y. The x-axis shows the probability of $P(Y = y_1)$.

indicates the number of binary questions that can be answered about Y from observing X_j. The definition of mutual information is based on the joint distribution of the variables in question:

$$I(X_j, Y) = \sum_{k=1}^{m_j} \sum_{i=1}^{n} P(x_{jk}, y_i) \cdot \log_2 \frac{P(x_{jk}, y_i)}{P(x_{jk}) \cdot P(y_i)}$$

The optimal test is thus the test X^* that is expected to yield the highest gain of information:

$$X^* = arg \max_X I(X, Y).$$

3.1.2 Gini Index

A second measure for test selection is called Gini index. It was introduced in 1912 by the statistician Corrado Gini as a measure of inequality or dispersion [39]. The *Gini index* of the probability distribution over Y is defined as:

$$G(Y) = 1 - \sum_{i=1}^{n} P(y_i)^2$$

For a binary variable of interest the Gini index is illustrated in Figure 3.1.

Similarly to the conditional entropy, the conditional Gini index after performing test X_j and yielding the test outcome x_{jk} is defined as:

$$G(Y|X_j = x_{jk}) = 1 - \sum_{i=1}^{n} P(y_i|X_j = x_{jk})^2$$

The expected Gini index after performing test X_j is defined as:

$$G(Y|X_j) = \sum_{k=1}^{m_j} G(Y|X_j = x_{jk}) \cdot P(x_{jk})$$

The optimal test is here the test that maximizes the Gini gain $G(Y) - G(Y|X_j)$.

Sent and van der Gaag [80] investigated the behavior of these two information measures for test selection. To illustrate the differences of the measures Figure 3.1 shows the Shannon entropy and the Gini index for a binary variable of interest Y. The x-axis shows the probability of $P(Y = y_1)$. We see that the Shannon entropy reaches higher values than the Gini index. However, for the purpose of test selection we are not so much interested in the precise value of the measures. We are rather interested in how the measures value a shift in the distribution of Y that is induced by the observation of a test result. For a binary variable of interest Y, Sent and van der Gaag observe that for probability distributions $P(Y = y_1) \in [0.37, 0.63]$ a shift is valued similarly by the Shannon entropy and the Gini index. However, for more extreme distributions the Shannon entropy values a shift higher than the Gini index. Thus, the measures are expected to occasionally select different tests. Nevertheless, the authors conclude that despite the possible differences in behavior both, Shannon entropy and Gini index are both well suited measures for test selection.

3.1.3 Incorporation of Costs

Glasziou and Hilden [40] criticize the above described approaches to test selection since the approaches disregard any type of costs of a test. They discuss how costs can be incorporated and introduce four test selection measures that handle costs. For the evaluation of the test selection measures, they choose five criteria:

1. If two tests are equal in information content, the selection measure should prefer that with the lower cost.

2. If the costs of two tests are equal, select the more informative one.

3. As the cost of a test with positive information content tends to zero, the selection measure should tend to a universal maximum. In other words, if a test is free of cost (neither time, money, or risks), it should always be performed before using more costly tests.

4. If the cost of a test exceeds the expected value of perfect information, the selection measure should indicate that the test is not worth doing.

5. Do not distinguish between diseases if they have identical conditional prognoses.

Both, information gain and Gini gain only satisfy the second criterion since they disregard costs. Glasziou and Hilden present four new test selection measures that attempt to satisfy more of these criteria. Each measure has its benefits and drawbacks. Therefore, the authors conclude that there is no optimal measure for all situations and recommend to rather choose an appropriate measure depending on the given situation.

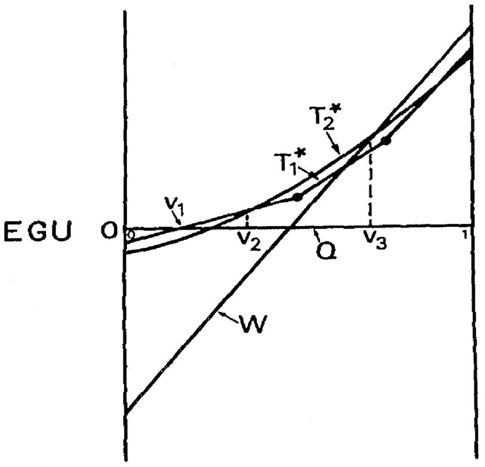


Figure 3.2: Graphical solution for test selection. The figure shows a sample solution for two tests T_1 and T_2. The graphs of the expected gain in utility (EGU) of Q, W, T_1^* and T_2^* are drawn together. For each value p of the prior probability distribution, the best approach is the one whose graph is highest at p. v_1, v_2, v_3 are the values of the prior probability at which the best approach changes. [33]

3.1.4 Incorporation of Utility

Further approaches to test selection discuss the definition of an utility function, that weights the expected benefits of an test against its costs and risks.

Doubilet [33] offers a mathematical approach to test selection for settings in which the presence or absence of a single disease is in question. The approach addresses three questions:

1. Should a test be done?
2. Which test (if any) should be performed if two or more are available?
3. If a test that can take on more than two values is performed, what is the correct cutoff point that determines how to proceed?

For each situation Doubilet considers two possible actions: W, a workup for the disease (treatment), or Q, do nothing or a less invasive treatment to W. Furthermore, it is assumed that the prior probability of a patient having the disease is known. The prior probability can depend on the available background information about the patient, as demographical data or information on previous test results. Moreover, the approach requires the definition of the expected utility of each possible action. The expected utility of an action is introduced as a function of the prior probability p. The expected gain in utility (EGU) is the difference between the expected utility and the expected utility of doing Q. An example for the solution to the problem which test to choose among two available tests T_1 and T_2 is illustrated in Figure 3.2. The graphs of the EGU of Q, W, T_1^* and T_2^* are plotted over the prior probability distribution of having a disease. For each prior probability p the highest graph indicates the appropriate course of action. In the given example, Q is the optimal approach for $p \in [0, v_1]$, T_1 is the optimal test for $p \in [v_1, v_2]$, and so on.

Andreassen [7] suggests to use an utility function as well. His approach is based on causal probability networks (CPN) that model medical domains. The nodes of the CPN represent possible tests, test results, symptoms, status of disease, and possible therapies. Here, the CPN is expanded with a number of utility nodes, meaning that some nodes have utilities associated with their states. The utility function then calculates for each of these nodes its contribution to the total utility. Doing so, it is possible to model the total utility by specifying multiple utility nodes rather than one utility that is a function of many nodes. The latter one is more difficult to specify in most applications. Having defined an utility function, test selection is performed by simply selecting the test X^* that maximizes the expected utility:

$$X^* = arg \max_{X_i} U(X_i) - U(X_0), \tag{3.1}$$

where $U(X_0)$ is the expected utility of doing no test and $U(X_i)$ is the expected utility of doing test X_i. Andreassen's approach is adapted for scenarios where sufficient expert knowledge is available in order to define the CPN and specify its associated utilities.

In this thesis, we approach test selection in the context of breast cancer diagnosis. Chapter 4 shows how to make use of mutual information in a data-efficient way. Chapter 5 introduces a new test selection method based on subgroup discovery, the data mining technique that is introduced in the following Section.

3.2 Subgroup Discovery

The task of *subgroup discovery* (SD) has been defined by Klösgen [54] and Wrobel [91] as: "Given a population of individuals and a property of those individuals that we are interested in, find population subgroups that are statistically 'most interesting', for example, are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest."

Consider, for instance, a population described by general demographic attributes and a target variable (property/attribute of interest) representing a disease status (disease, non-disease). Then, an subgroup discovery algorithm might come up with a subgroup identified by a conjunction of two conditions ($age > 75 \wedge gender = female$). Here, the subgroup description consists of two attribute-value tests, and it selects a set of persons with a particularly high prevalence of the disease (85% compared to 43% in the entire population). Figure 3.3 presents a visualization of this subgroup as a pie chart.

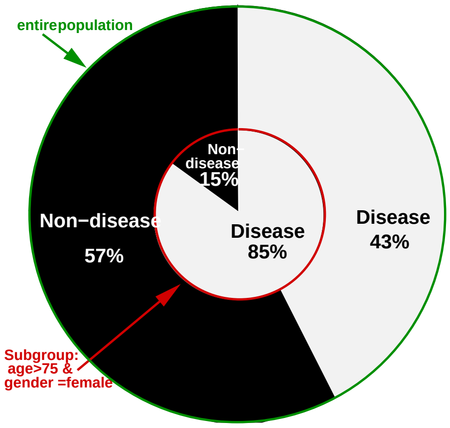


Figure 3.3: The nested cake diagram visualizes a subgroup S. The outer circle displays the distribution of classes (”Disease” and ”Non-Disease”) over the entire population. The inner circle displays their distribution over S. The radius of the inner circle represents the size of S (number of examples covered by S). S can be called an interesting subgroup, as its distribution differs strongly from the one of the entire population.

Subgroups represent a subset of individuals of a population. In contrast to clusters obtained by clustering algorithms (see Section 3.3), each subset is describable by a *subgroup description*, *e.g.,* a conjunction of attribute-value pairs selected from the features of the training set. In general, subgroups have the form $Class \leftarrow Cond$, where $Class$ is the property of interest and $Cond$ is the subgroup description. Thus, subgroups can also be interpreted as rules: if $Cond$ then $Class$.

Subgroup Discovery is considered as a hybrid of predictive and descriptive methods.

The discovered rules offer a descriptive survey of interesting patterns in the data. And if *Class* refers to the actual class label of the data, the rules can be interpreted as classification rules that enable a classification of unseen examples. However, subgroup discovery differs from classification rule learning because it aims for a descriptive exploration of the entire population. Classification rule learners, on the other hand, aim for an accurate classification. A further requirement for subgroups is that their description should be easily understandable by the users. Therefore, one tries to keep the description short.

3.2.1 Subgroup Discovery Algorithms

During the last years a variety of algorithms have been developed to solve this tasks. Most ideas originate from established predictive or descriptive methods that are modified to discover subgroups. Well known SD algorithms include EXPLORA [54], MIDOS [91], APRIORI-SD [53], CN2-SD [56], RSD [57], and SD-MAP [9]. For each algorithm one has to define:

1. the subgroup description language (*e.g.,* conjunctions of attribute-value pairs),
2. the target variable (*e.g.,* nominal or numerical),
3. the quality function (see Section 3.2.2), and
4. the search strategy (*e.g.,* exhaustive or with search heuristic).

MIDOS

Wrobel *et al.* [91] proposed MIDOS, an algorithm that can discover subgroups in multi-relational databases. MIDOS is a top-down algorithm that uses frequency pruning and an additional pruning by incorporating an optimistic quality estimate function.

APRIORI-SD

APRIORI-SD [53] is an adaptation of the association rule learning algorithm APRIORI [2] to subgroup discovery. APRIORI-SD starts with generating subgroups described by a single attribute-value pair. Subsequently, it generates subgroups with longer (and thus more specific) descriptions. Subgroups are only kept if they contain more examples than *minsupport*. All smaller subgroups are pruned, and no subgroups more specific than these are generated.

CN2-SD

CN2-SD [56] is a modification of the CN2 rule learner [25], which uses a covering heuristic to provide a descriptive exploration of the entire population. The difficulty of applying a common rule learner to this task is that its covering algorithm is not designed for finding subgroups. It does not take into account examples as soon as they are covered by a rule. This implies that rules discovered in a later iteration are built on a smaller and thus biased subset of examples. Therefore, only the first few rules found by a rule learner are appropriate for subgroup discovery, i.e., they have a sufficiently large coverage. To overcome this problem, CN2-SD assigns the weight $\frac{1}{i+1}$ to each example, where i is the number of rules (subgroup descriptions) covering the example. Initially, the weight of each example is set to one. Whenever a rule is found, the weight is decreased for each covered example.[1] To find large and interesting rules even in later iterations, CN2-SD uses the *weighted relative accuracy (WRAcc)* heuristic. For a subgroup with description $Cond$ and target variable $Class$, it is defined as:

$$WRAcc(Class \leftarrow Cond) = \frac{n'(Cond)}{N'} \cdot (\frac{n'(Class.Cond)}{n'(Cond)} - \frac{n'(Class)}{N}) \tag{3.2}$$

where N is the number of examples, $n'(Class)$ is the sum of weights of examples in $Class$, N' is the sum of weights of all examples, $n'(Cond)$ is the sum of weights of examples covered by the subgroup, and $n'(Class.Cond)$ is the sum of weights of examples that are covered by the subgroup and actually fall into the class. The algorithm iteratively adds the rule with the highest $WRAcc$ measure. $WRAcc$ tends to find rules for examples that are least frequently covered by previously discovered rules. Furthermore, $WRAcc$ ensures a balance between generality and relative accuracy. This results in shorter rules which are thus easier to comprehend compared to the outcome of a rule induction algorithm.

RSD

The RSD (relational subgroup discovery) algorithm [57] is very similar to CN2-SD algorithm. It differs from CN2-SD in the fact that it can discover subgroups in a relational dataset. This is achieved with a propositionalization step that generates first-order features. Representing the data through these features enables the application of the CN2-SD algorithm. A further improvement of RSD is its *modified weighted relative accuracy (mWRAcc)* heuristic

$$mWRAcc(Class \leftarrow Cond) = \frac{n'(Cond)}{N'} \cdot (\frac{n'(Class.Cond)}{n'(Cond)} - \frac{n(Class)}{N}) \tag{3.3}$$

which computes the prior probability computed with $\frac{n(Class)}{N}$ instead of $\frac{n'(Class)}{N}$, where $n(Class)$ is the number of examples in $Class$. This leaves the prior probability unaffected by modifications of example weights. RSD also introduces a further pruning step that allows to stop specifying subgroups whose covered examples fall into the same class. Furthermore, subgroups with an $mWRAcc$ below a user-defined threshold are pruned, because no specialization can yield a higher $mWRAcc$ measure.

SD-MAP

In 2006, Atzmüller *et al.* introduced the algorithm SD-MAP [9]. While the previously introduced approaches feature a heuristic search, SD-MAP features an exhaustive search,

[1] Instead of the additive weight, CN2-SD also works with a multiplicative weight γ^i with $0 < \gamma < 1$. The parameter γ has to be set in advance. Its default is 0.9.

which guarantees the identification of all interesting subgroups contained in the data. This can be efficiently implemented by using the FP-growth algorithm, which is an efficient algorithm for association rule mining developed by Han *et al.* [45]. Furthermore, the use of an FP-tree as the underlying data structure enables the computation of the subgroup quality immediatly during the generation process of the frequent patterns. Moreover, SD-MAP can handle missing values and disjunctions of attribute-value pairs in the subgroup description. Atzmüller *et al.* showed SD-MAP to be faster than APRIORI-SD.

Further Variants of SD Algorithms

Most SD algorithms are designed for nominal attributes. This means that numerical attributes have to be discretized which can yield suboptimal results. Grosskreutz *et al.* [43] investigate how this can be circumvented by allowing numerical values in the subgroup description. The handling of numerical target variables is discussed by Klösgen *et al.* [54] and Atzmüller *et al.* [8].

Whereas many existing SD algorithms quite often only offer pruning according to frequency [53], in recent years more and more other pruning constraints are investigated, for instance, optimistic estimate pruning [91, 42]). In Chapter 5 we will introduce the algorithm $SD4TS$ that enables pruning of subgroups that can not yield a reasonable quality.

In order to avoid the discovery of redundant subgroups, Boley *et al.* [17] proposed to search for equivalence classes of descriptions. These equivalence classes have unique maximal representatives that form a closure system. They show that the search space and the number of output subgroups can be significantly reduced by using the concept of equivalence classes. For the same problem Atzmueller *et al.* [10] proposed to do a clustering of subgroups based on their similarity regarding the set of instances they cover. Similarity is measured by the Jaccard index. Two subgroups are merged if their similarity is above a user-defined threshold. This enables a surpressing of irrelevant subgroups which saves time for the user that has to assess the resulting subgroups.

3.2.2 Interestingness of a Subgroup

Several criteria can be considered for the evaluation of an individual subgroup of the form $Class \leftarrow Cond$ in a population of the size N. These include size, support, coverage, and significance. For the denotion of parameters refer to Table 3.1 that summarizes the denotion of several authors.

- The *size* of a subgroup is simply the number of examples $n(Cond)$ that are covered by the subgroup.
- The *coverage* of a subgroup is the percentage of examples covered: $\frac{n(Cond)}{N}$.
- The *support* of a subgroup is the percentage of examples covered by the subgroup that actually share the $Class$ value: $\frac{n(Class.Cond)}{N}$.

SD-MAP	CN2-SD, RSD	Meaning
N	N	size of population
n	$n(Cond)$	size of subgroup
TP	$n(Class)$	size of target class
FP	$N - n(Class)$	size of non-target class
tp	$n(Class.Cond)$	number of true positive examples
fp	$n(Cond) - n(Class.Cond)$	number of false positive examples
p_0	$\frac{n(Class)}{N}$	relative frequency of the target class in the population
p	$\frac{n(Class.Cond)}{n(Cond)}$	relative frequency of the target class in the subgroup

Table 3.1: Denotion of subgroup parameters in different subgroup discovery algorithms.

- To determine the *significance* of a subgroup, we use the likelihood ratio [51]:

$$Sig(Class \leftarrow Cond) = 2 \cdot \sum_{j=1}^{k} n(Class_j.Cond) \cdot log \frac{n(Class_j.Cond)}{n(Class_j) \cdot p(Cond)} \tag{3.4}$$

  It shows how significantly different the class distribution in a subgroup is from the prior class distribution by incorporating both the size of a subgroup and the unusualness of the subgroup given its distribution of the target attribute. The significance can be used to estimate the p-value, which indicates the statistical significance of the rule. It is calculated from the χ^2-distribution with $k-1$ degrees of freedom (k=number of classes). As usual, a subgroup is considered significant, if its p-value is below 0.05.

- For binary target variables, the following quality function aims for possible large subgroups with an unusual target distribution.

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N-n}} \tag{3.5}$$

- The following quality function evaluates again the distributional unusualness of a subgroup regarding the binary target variable. q_{RG} measures only the relative gain between the target share of the subgroup and the target share of the total population. It hence prefers subgroups that are different from the original target distribution.

$$q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)} \tag{3.6}$$

3.2.3 Related Approaches

Kralj *et al.* [73] intoduce *supervised descriptive rule discovery* as a framework that comprises *subgroup mining* as well as *contrast set mining*, and *emerging pattern mining*. These are three areas that basically address the same task, although they use a different terminology. Kralj *et al.* thus show that algorithms, heuristics, and visualizations developed for one of the areas can be translated to solve tasks of the other two areas. They also give a survey of different visualization techniques for the discovered rules.

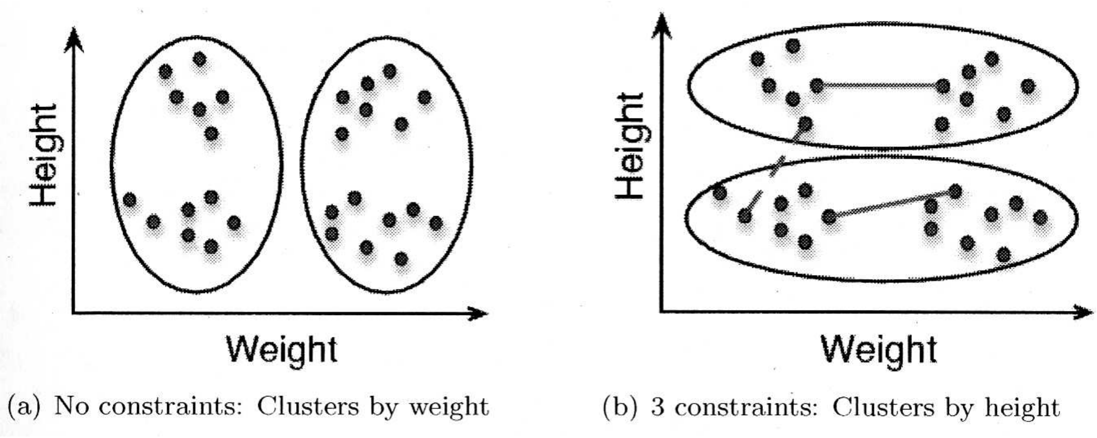
A further approach related to subgroup discovery is *exceptional model mining* [58], which enables the discovery of subgroups with a more complex target concept: a model and its fitting to a subgroup. It performs a level-wise beam search and explores the best t subgroups of each level.

The last related approach, we are specifying here, is *bump hunting* or *patient rule induction (PRIM)* introduced by Friedman and Fisher [36]. The aim of bump hunting is the identification of subgroups with a high or low average value of a given target variable. Moreover, these subgroups should be describable by rules on the patient attributes.

3.3 Constrained Clustering

The third data mining technique we investigate in this thesis is *constrained clustering*. Clustering is one of the most popular data mining technique since it enables the user to get an overview of the structure of the data and to identify major patterns or trends without any supervisory information such as class labels. Thus, it belongs to the unsupervised data mining methods. In general, clustering provides a partitioning of a data set into groups (clusters), in such a way that data points within the same group are as similar as possible to each other and as different as possible to data points from other groups. These groups can be interpreted as meaningful subpopulations, similar to subgroups. Of course, one of the main prerequisites of an effective clustering is an adapted definition of similarity. This has to be achieved by defining a similarity or distance function. This function computes for each pair of data points of the given domain how similar they are to each other based on their attributes. One of the most commonly used distances is the Euclidean distance. A discussion of further distance functions can be found in an article by Jain *et al.* [49].

Moreover, clustering can be considered as a search for an optimal solution. Since there are various opinions about optimality, various clustering algorithms have been developed, each pursuing different search heuristics or different scoring functions, which define the quality of a given clustering. Kaufman and Rousseeuw [52] and Berkhin [15] describe the majority of these algorithms in detail. The variety of clustering algorithms can be divided into *partion-based* and *hierarchical* algorithms. Partition-based algorithms start with a specified number k of clusters the clustering has to consist of. The resulting clustering is then the definition of the k clusters. The most prominent example for these algorithms is k-means [52]. Hierarchical clusterings, in contrast, provide a complete cluster structure, which enables the user to compare solutions for different k. The results of a hierarchical clustering is often graphically displayed as a dendrogram. Hierarchical clustering algorithms can be divided into *agglomerative* and *divisive* algorithms. Agglomerative algorithms start form a clustering where each cluster constains exactly one data point. Then, the algorithm iteratively merges the closest pair of clusters until the clustering consists one single cluster containing all data points. Agglomerative algorithms vary in the definition of distance of clusters. Prominent examples are single-linkage, complete-linkage, average-linkage, or Ward's method of minimum variance. Divisive algorithms work contrariwise and start with a clustering consisting of one cluster that contains all data points. In each step of the algorithm, the most heterogeneous cluster is divided into two clusters until the clustering consists of clusters that contain only one data point. Prominent examples for divisive clustering algorithms are DIANA [52] or Bisecting-k-means [84]. Further variants of clustering include probabilistic clustering, where the result is not a definite assignment of each point to a cluster but rather probabilties for a point to belong to the clusters, and density based clustering like DBSCAN [35].



(a) No constraints: Clusters by weight

(b) 3 constraints: Clusters by height

Figure 3.4: Example: Clustering (k=2) with hard pairwise constraints. Must-link constraints are indicated with solid lines, and cannot-link constraints are indicated with dashed lines [13].

3.3.1 Incorporation of Constraints

In many settings, the user has domain knowledge or additional information about the data and thus previous knowledge about the possible clusters that can be retrieved from the data. This domain knowledge or user preferences can be incorporated into the process of clustering in the form of constraints. In this case we speak of *constrained clustering* which belongs to the field of constraint-based mining [19]. The subject of constrained clustering has been addressed by many researchers in recent years. The current state of the art is thoroughly presented in the book “Constrained Clustering: Advances in Algorithms, Theory, and Applications“ by Basu *et al.* [13].

Basically, there are three types of constraints: constraints on the instances, constraints on the clusters, and constraints on the clusterings. *Instance-level constraints* are pairwise constraints on cluster membership. These are divided into *must-link constraints*, meaning that two particular instances have to be put into the same clusters, and *cannot-link constraints*, meaning that two particular instances have to be put into two different clusters. Figure 3.4 illustrates an example of two-dimensional data points. In this example, there are two reasonable ways of partitioning the data into two groups: by weight (a) or by height (b). An unsupervised clustering method will result in one of them. However, if the user is only interested in clusters by height, then she can express this by specifying constraints as the two must-link constraints between pairs of points with similar height and a cannot-link constraint between two points with different heights. This results in the clustering in Figure 3.4 (b).

The second type of constraints refers to the clusters that can be combined into a clustering. These constraints can restrict the size of the clusters, e.g., *minimum cluster size constraints* or *maximum cluster size constraints*, which can be adopted in order to avoid solutions with empty or singleton clusters [32]. Moreover, it is possible to formulate *balancing constraints* which yield clusterings consisting of clusters of comparable size [11]. Constraints about the clusters can also relate to their descriptiveness. *Itemset constraints* [82], for instance, require that each cluster can be described by an itemset, i.e., a conjunction of attribute-value pairs of the instances contained in the cluster.

The third type of constraints are constraints on entire clusterings: *Set-level constraints* control the clusters that can go into a clustering depending on other clusters, for instance, certain cluster can require or exclude other clusters. *Clustering constraints* determine the form of a clustering, for instance, whether the clusters are allowed to overlap or even encompass other clusters and whether the clustering has to be complete. *Optimization constraints* determine the scoring function to evaluate the quality of a clustering. In this thesis we address the latter type of constraints and show how they can be translated into linear constraints. This enables us to find an optimal clustering by solving integer linear problems (see Chapter 8).

3.3.2 Constrained Clustering Algorithms

One of the first algorithms developed for constrained clustering is *COP-KMEANS* [87], proposed by Wagstaff *et al.* in 2001. This approach incorporates pairwise constraints into k-means. Similar to k-means, COP-KMEANS starts with k initial cluster centers $\mu_1, \ldots, \mu_k$. In a second step, each instance x is assigned to the cluster i with the minimal distance between x and μ_i. However, in contrast to k-means in COP-KMEANS each instance is assigned to a cluster only if it dos not violate any of the given pairwise constraints. Otherwise it is assigned to the next closest cluster if no constraint is violated by the assignment, and so on. The third step of COP-KMEANS is again similar to k-means the update of cluster centers followed by the iteration of step two and step three until the assignment remains stable. A drawback of COP-KMEANS is that it may fail to find a satisfying solution, even when there exists one. This is caused by the greedy way of assigning instances without allowing any type of backtracking. In this approach constraints are considered to be *hard* constraints, meaning that each solution has to satisfy all constraints. Since domain knowledge is often more heuristic than exact, it is often useful to consider constraints as *soft* constrains, meaning that a solution is possible that satisfy only a subset of the constraints. Algorithms on soft constraints thus aim to satisfy as many constraints as possible.

In order to handle soft constraints, Davidson and Ravi [30] introduce a new derivative of k-means. They introduce a new error function for the cluster assignment: the *constrained vector quantization error* (CVQE). This error is the sum of the constrained vector quantization errors of the individual clusters. For a cluster, this error consists of three parts: the sum of the distances of each instance in the cluster to the cluster center (regular distortion), a term for all violated must-link constraints and a term for all violated cannot-link constraints. If a must-link constraint is violated then the cost is equal to the distance between the cluster centroids containing the two instances that should have been in the same cluster. Similarly, if a cannot-link constraint is violated the cost is the distance between the cluster centroid both instances are in and the nearest cluster centroid to one of the instances. Both violation costs are in units of distance, as

$$
\begin{array}{|llll|}
\hline
\underset{T}{\text{minimize}} & & \sum_{i=1}^{n} \sum_{h=1}^{k} T_{i,h} \cdot (\frac{1}{2}||x_i - \mu_h||^2) & \\
\hline
\text{subject to} & \text{(i)} & \sum_{i=1}^{n} T_{i,h} \geq \tau_h & \forall h \in \{1, \ldots, k\} \\
 & \text{(ii)} & \sum_{h=1}^{k} T_{i,h} = 1 & \forall i \in \{1, \ldots, n\} \\
 & \text{(iii)} & T_{i,h} \geq 0 & \forall i \in \{1, \ldots, n\} \quad h \in \{1, \ldots, k\} \\
\hline
\end{array}
$$

Table 3.2: Cluster Assignment Step of Constrained k-means by Demiriz *et al.* [32]. $T_{i,h} = 1$ if the data point x_i is closest to the cluster center μ_h, and it is 0 otherwise.

is the regular distortion.

In the first step of the constrained k-means algorithm the CVQE is minimized. This is achieved by assigning instances such that they minimize the introduced error term. For instances that are not part of constraints, this involves as in regular k-means, assigning the instance to the cluster with the closest cluster center. For pairs of instances in a constraint, for each possible combination of cluster assignments, the CVQE is calculated and the instances are assigned to the clusters that minimally increases the CVQE. In the second step the cluster centers are updated. This is done in such a way that if a must-link constraint is violated, the cluster centroid is moved towards the other cluster containing the other point. If a cannot-link constraint is violated the cluster centroid containing both constrained instances should be moved to the nearest cluster centroid so that one of the instances eventually gets assigned to it, thereby satisfying the constraint. With this algorithm constraints are satisfied unless it is less costly to violate it, *e.g.,* if two instances in a must-link constraint are very distant according to the distance function.

Demiriz *et al.* [32] tackle constraints on the clusters, in particular, the constraint on a minimum cluster size. Their approach is presented as a linear optimization problem. The objective is similar to the objective of k-means to minimize the sum of the distances of each instance x_i to the cluster center μ_h of the cluster h it is assigned to. They introduce k additional constraints requiring that each cluster h has at least a minimum number of τ_h points (see constraint (i) in Table 3.2). Note that the constraints (ii) and (iii) ensure that each data point is assigned to exactly one cluster. The optimization is divided into two steps. The first step optimizes the cluster assignment of the instances where the k cluster centers remain fix (see Table 3.2). The second step is the update of the k cluster centers where the cluster assignment of the instances to the clusters remains fix. The two steps are iterated until no cluster centers are updated anymore in the update step.

In Chapter 8 we tackle constraints on clusterings. We provide a framework that shows how these constraints can be translated into linear programs and then be solved by optimization packages. We show how this method performs on two medical data sets.

4 Test Selection

Many real-world applications, where data gathering and analysis play a central role, are faced with the fundamental question of how to select variables (to be instantiated or measured) that can efficiently reduce uncertainty. Here, we use the concept of conditional mutual information to approach problems involving the selection of variable observations in the area of medical diagnosis. Computing mutual information requires estimates of joint distributions over collections of variables. However, in general computing accurate joint distributions conditioned on a large set of variables is expensive in terms of data and computing power. Therefore, one must seek alternative ways to calculate the relevant quantities and still use all the available observations. We describe and compare a basic approach consisting of averaging mutual information estimates conditioned on individual observations and another approach where it is possible to condition on all observations at once by making some conditional independence assumptions (while the first approach does not require making any data modeling assumptions). This yields a data-efficient variant of information maximization for test selection. We present experimental results on public heart disease data (our test domain) and data from a controlled study in the area of breast cancer diagnosis (our main application).

4.1 Motivation

In this chapter, we present an information-theoretic solution to the problem, which is able to take into account background knowledge about patients, and gives relatively accurate results without requiring excessively large datasets. We assume a number of i.i.d. data points that are observed partially or fully (some components or dimensions are observed while others are not). We would like to use these observations, so that for a given data point, we are able to determine which components should be observed or queried to gain the most information about a variable of interest (*e.g.,* class label).

In this setting, joint probability estimates are built from the available data. When a new partially observed data point is given, we use the above estimates to determine what unobserved portion should be tested or queried to gain the most information about the quantity of interest. Two common problems with information-based methods are that they require computing joint distributions among a large number of variables and conditional distributions conditioned also on many variables. This, in general, creates data requirements that are difficult to satisfy in standard real world applications (more data is required in order to build accurate estimates to probability functions). Using very limited amounts of data, we experimentally show how simple conditional independence assumptions can be used to address the problem at hand.

The chapter is organized as follows. Section 4.2 introduces our application on breast cancer diagnosis. In Section 4.3 we describe the general problem of test selection and in Section 4.4 we propose two approaches to solve this problem. We tested and applied the approach on data from a study on breast cancer diagnosis and, as a test domain, on heart disease data from the UCI machine learning repository and present the empirical results in Section 4.5. Section 4.6 surveys related work.

4.2 Application: Breast Cancer Screening and Diagnosis

In Section 2.1.1 we introduced imaging modalities that are used for breast cancer diagnosis. They include Film Mammography, Digital Mammography, Ultrasound, and Magnetic Resonance Imaging. When a specific patient is under scrutiny for breast cancer, it is not clear which of these modalities is best suited to answer the basic question to whether the patient has or does not have cancer. Deciding which modality to use for a particular case, usually requires considerable experience of the health care workers [47]. For this particular problem we use the dataset collected in the HUP study which is described in Section 2.1.

Assuming we would like the patient to be exposed to the least number of examinations (modalities in this case), we now want to find a way to determine for each lesion the modality that provides the most information. In this context, our system automatically learns from previous patient records and at the time of diagnosis, taking into account the specifics of the patient in consideration, it suggests one or several suitable (clinical) courses of actions to efficiently arrive at a diagnosis with as high certainty as possible. The method works by iteratively reducing the uncertainty about the variable of interest (in our example, the occurrence of cancer) by making locally optimal suggestions of what variable to investigate (observe) next.

4.3 Method

Consider a collection of patient records containing data generally indicative of various clinical aspects associated to the patient, *e.g.,* patient demographics, reported symptoms, results from laboratory or other tests, and patient disease/condition. Let us define each single element (source) of patient data as a random variable V_m, *e.g.,* patient age (demographics), presence of headache (symptom), blood pressure (test result), or occurrence of diabetes (disease).

In this work, we consider the set of variables $V = \{V_1, \ldots, V_M\}$ and we assume each variable to be discrete with a finite domain. For a given patient, a few of these variables may have been observed while others may have not. In some cases, we know such values of the variables with some probability.

Let us now consider the time when a diagnosis needs to be made for a specific patient, for which some of these variables have been observed, denoted as background attributes $Z_i \in V$, for $i \in \{1, ..., k\}$; *e.g.,* demographic information and patient symptoms. The remaining $M - k$ variables denoted X_j have not been observed yet; *e.g.,* the result of a lab test. Finally, for the patient, let Y denote a variable in which we are ultimately interested, but that we have not observed, such as the occurrence of cancer. This variable may not be observable directly or it may be observable at a high cost. Thus, we prefer to rely on other sources of information and try to infer the value of this variable.

Preliminary analysis of the data presented in Chapter 2.1 showed that the four modalities perform differently and none of them performs perfectly for all lesions (Section 2.1.2). Thus, we now want to find a way to determine for each lesion the modality that provides the best performance. Of course, there can be several modalities that perform equally well for a lesion and there can be lesions that are assessed poorly by all modalities.

The goal is now to take into consideration all information available about a patient, and then determine the optimal modality. We focus on a solution by applying mutual information to determine the modality that maximizes the information about the condition of the patient (biopsy result).

4.3.1 Formalization

For the introduced application, we have the following variables.

- Y = Biopsy result; Domain(Y) = {malignant, atypia benign, benign}
- X_j = Tests $j \in$ {DMAM, FMAM, MRI, USND};
  Domain(X_j) = $\{0, 1, \ldots, 5\}$ (BI-RADS Assessment of Test)
- $Z_1, \ldots, Z_k$ = Attributes about patients and lesions. Discrete values and different domain sizes.

4.3.2 Basic Methodology

Formally we are interested in optimizing,

$$X^* = \texttt{arg}\max_j I(X_j, Y | Z_1 = z_1, \ldots, Z_k = z_k) \tag{4.1}$$

where the quantity I is the mutual information between our variable of interest Y and the variables whose values we could potentially obtain X_j, conditioned on the fact that we already know $Z_1 = z_1, \ldots, Z_k = z_k$. Using the definition of mutual information (see Section 3.1.1) and the shorthand $\mathbf{z} = (z_1, \ldots, z_k)$ to denote the assignment of multiple variables we have that the function of interest is:

$$I(X_j, Y|\mathbf{z}) = \sum_{x_j} \sum_{y} P(x_j, y|\mathbf{z}) \cdot \log \frac{P(x_j, y|\mathbf{z})}{P(x_j|\mathbf{z}) \cdot P(y|\mathbf{z})} \tag{4.2}$$

The set of variables X_j is finite. Assuming the variables follow a multinomial joint probability distribution that can be reliably estimated, this problem can be solved by testing each of the potential candidate variables individually to see which provides the most information.

In order to arrive at a diagnosis with high certainty, one observation may not be enough. In some cases we may be allowed to observe more than one variable. Thus, this process can be repeated iteratively until a user-defined precision or confidence level is achieved. For example, this limit can be defined in terms of the amount of information that is left in the variable of interest (entropy).

Once a variable has been tested and observed, it can be incorporated as part of the background knowledge. Specifically, let us say that X_i was observed, we can now solve the following refined problem:

$$\mathtt{arg}\max_{j\neq i} I(X_j, Y|x_i, \mathbf{z}) \tag{4.3}$$

that is the maximization problem needs to be updated once a test (generating an observation) has been performed. Another extension of this approach includes optimizing over two variables at once.

Ideally, the quantity to optimize is the mutual information conditioned on all available data or observed variables . However, in practice this may be difficult because in general the conditional joint probabilities $P(x_j, y|z_1, z_2, \dots, z_k)$ cannot be properly estimated from limited data for large k. The data required to properly estimate the conditional joints may grow exponentially with k. In addition, the number of unobserved variables plays an important role in terms of computational complexity (which can also grow exponentially with the number of unobserved variables).

4.3.3 Feature Selection

We consider only significant background attributes Z_i. To determine the significance of an attribute X for another attribute Y we calculate $signif(X, Y) = I(X, Y) - \frac{(|X|-1)\cdot(|Y|-1)}{\#\mathtt{lesions}}$. An attribute X is significant if $signif(X, Y) > 0$, otherwise it is discarded from our calculation. Table 4.1 shows the available attributes X, their information gain about the biopsy result Y, and their $signif$-value. The table also displays for how many lesions no attribute value is known (missing value). The highest information gain results for the attributes that represent a direct test result. Of course, these attributes are unknown before the examination is done. Therefore, we can not include them into our calculation. We therefore select the fifteen other attributes with positive $signif$-value as $\mathbf{z}$; these are the attributes between the two horizontal lines in the table.

4.4 Combining all Available Background Information

Since with limited data our estimate of $P(x_j, y|z_1, z_2, \dots, z_k)$ will not be accurate, we seek a *data-efficient* method (one whose data requirements do not grow exponentially with k) to compute or approximate the above quantity and thus be able to use all available background information to make decisions over what test should be chosen.

X	I(X,Y)	$signif(X,Y)$	#missing values
Biopsy result	1.160	1.140	0 (0.0 %)
CLF BI-RADS	0.369	0.325	0 (0.0 %)
USND BI-RADS	0.286	0.242	29 (12.8 %)
MRI BI-RADS	0.284	0.240	16 (7.0 %)
FMAM BI-RADS	0.259	0.214	48 (21.1 %)
DMAM BI-RADS	0.228	0.183	30 (13.2 %)
Gail Model Lifetime Risk	0.133	0.116	180 (79.2 %)
Gail Model 5 Year Risk	0.121	0.103	180 (79.2 %)
has child	0.093	0.084	180 (79.2 %)
Claus Model lifetime risk at age 79	0.086	0.068	180 (79.2 %)
age at menarche	0.094	0.068	180 (79.2 %)
calcification	0.060	0.051	68 (29.9 %)
has diagnositc MAM	0.024	0.015	12 (5.3 %)
1st degree relative with BC	0.021	0.012	8 (3.5 %)
age	0.027	0.010	8 (3.5 %)
has last MAM	0.014	0.005	8 (3.5 %)
Educational level	0.030	0.003	9 (4.0 %)
currently use of oral contraceptives	0.011	0.002	9 (4.0 %)
ethnic group	0.019	0.002	8 (3.5 %)
has breast USND	0.010	0.001	11 (4.8 %)
menopausal status	0.018	0.000	9 (4.0 %)
has suspicious MAM	0.008	-0.001	11 (4.8 %)
relatives with bc	0.008	-0.001	8 (3.5 %)
has breast MRI	0.006	-0.003	10 (4.4 %)
breast density	0.005	-0.004	9 (4.0 %)
number of 2nd degree relatives	0.005	-0.004	5 (2.2 %)
curr. estrogen repl. or oral contrac.	0.003	-0.006	18 (7.9 %)
curr. estrogen repl. therapy	0.003	-0.006	8 (3.5 %)
Annual family income	0.003	-0.006	39 (17.2 %)
marital status	0.003	-0.015	9 (4.0 %)
Insurance plan	0.003	-0.015	8 (3.5 %)

Table 4.1: How much information provides a single attribute X about the biopsy result?

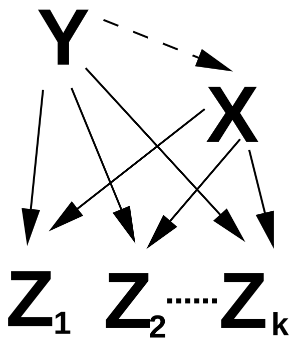


Figure 4.1: Bayes network: Background attributes Z_i depend only on the test result of Test X_j and the Biopsy outcome Y

4.4.1 Information Averaging

A simple heuristic consists on averaging the information conditioned on each background variable Z_i separately,

$$I(X_j, Y|\mathbf{z}) \approx \frac{1}{k} \cdot \sum_{i=1}^{k} \sum_{z_i} \alpha(z_i) I(X_j, Y|z_i), \tag{4.4}$$

where $\alpha(z_i)$ is equal to the prior $P(z_i)$ if z_i is not observed and $\alpha(z_i) = 1(z_i)$ if observed[1]. We will use this as our baseline method.

One way to think about this heuristic is to consider k independent models involving Y, X_j, Z_i of the form $P(y, x_j, z_i) = P(z_i|x_j, y)P(x_j, y)$. Mutual information can be computed for each model separately given the observed data. Similar to model averaging, the mutual information of X_j about Y can be found by averaging the individual information values, giving rise to the above equation.

4.4.2 Exploiting Conditional Independence Assumptions

We have so far assumed a very general model of the data, specifically a full multinomial distribution with a large number of degrees of freedom. However, if we relax this assumption, it turns out that we can find a family of models for which computing the mutual information of interest is computationally and data-efficient.

Our general approach is to find good models that make stronger conditional independence assumptions (simpler joint models), so that when the z_i's are observed, the computation of $I(X_j, Y|\mathbf{z})$ does not have large data requirements.

We consider the Bayes net in Figure 4.1, where the background information Z_i depends on the test result X_j and the biopsy outcome Y. This network should not be looked at as depicting causality, but simply as a statistical model. We obtain a new probability

[1] $1(z_i)$ is equal to one if z_i is the observed value for Z_i and zero otherwise

distribution Q:

$$Q(x_j, y, \mathbf{z}) = P(x_j, y) \cdot \prod_{i=1}^{k} P(Z_i = z_i | x_j, y). \tag{4.5}$$

When using this model, we obtain the following expression for the conditional mutual information:

$$I(X_j, Y|\mathbf{z}) \;=\; \sum_{x \in X_j} \sum_{y \in Y} Q(x, y|\mathbf{z}) \log \frac{Q(x, y|\mathbf{z})}{Q(x|\mathbf{z}) \cdot Q(y|\mathbf{z})} \tag{4.6}$$

The above Bayes network defines another multinomial model. This has the particular advantage that it only requires computing joint distributions of at most three variables for the mutual information computations above. This holds even if we do not know what variables Z_i will be observed beforehand which is beneficial since we indeed do not know what variables will be observed for a particular case (patient).

Learning and Inference in the New Model

For learning the model, we are interested in finding $P(z_i|x_j, y) \forall i \in \{1, ..., k\}$ and $P(x_j, y)$ since these distributions fully define the model. Using the maximum likelihood criterion, we can see that:

$$P(z_i|x_j, y) \;=\; \frac{count(Z_i = z_i, X_j = x_j, Y = y)}{count(X_j = x_j, Y = y)} \tag{4.7}$$

$$P(x_j, y) \;=\; \frac{count(X_j = x_j, Y = y)}{count(X_j = any, Y = any)}. \tag{4.8}$$

Inference can be performed efficiently also. Denoting Z_o as the background variables observed, and Z_u those unobserved[2], we have the following expression for the posterior probability over the domain of Y:

$$P(y|x_j, z_o) = \frac{\sum_{z_u} P(x_j, y) \prod_i P(z_i|x_j, y)}{\sum_{z_u} \sum_y P(x_j, y) \prod_i P(z_i|x_j, y)}, \tag{4.9}$$

where i indexes all (observed and unobserved) variables. This expression can be easily obtained from the definition of the joint distribution implied by the model.

In summary, we have found a method that allows us to use all the available background information for deciding what test to perform. For this we have relaxed the model assumptions and use a class of models whose data requirements are more practical. Inference and learning are computationally efficient in these models as well.

[2] So far we have assumed that all of the Z_i variables are observed since they are considered background information. A more general assumption can be made by letting some of the background variables be unobserved.

	Averaging	Joint Model (Q)
$P(y\|x_j, \mathbf{z})$	$\frac{1}{k} \cdot \sum_{i=1}^{k} P(y\|x_j, z_i)$	$\frac{P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}{\sum_{y\in Y} P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}$
$H(Y\|x_j, \mathbf{z})$	$\frac{1}{k} \sum_{i=1}^{k} H_P(Y\|x_j, z_i)$	$H_Q(Y\|x_j, \mathbf{z})$

Table 4.2: Approximations of the probability distribution and entropy given k background attributes ($\mathbf{z} = (z_1, \ldots, z_k)$)

4.5 Validation and Results

First, we test our approaches on public heart disease data consisting of 920 patients and 14 attributes (two background attributes, 11 observable test attributes, and one binary class attribute).[3] On this dataset each patient is an instance. The variable of interest Y expresses if the patient has heart disease (411 patients) or not (509 patients). In a second step, we perform experiments on the breast cancer diagnosis data introduced in Chapter 2.1 (15 background attributes, 4 observable test attributes, and one binary class attribute). For our experiments on the breast cancer data, we use lesions as instances. We only consider lesions that were subject to biopsy: from 132 patients we analyze 216 biopsied lesions, 134 malignant and 82 benign. We use the biopsy result as the variable of interest Y. Note that the number of data points is small given the dimensionality of the data, which is quite common in controlled medical studies where the cost of data gathering is usually high.

We validate the two approaches considered above for each instance in a leave-one-out validation.[4] First, we determine for each instance which test X^* maximizes the information given the patient specific attributes $\mathbf{z}$. After observing the selected test $X^* = x_j$, we decide for the most likely diagnosis $y^* = arg \max_{y \in Y} P(y|x_j, \mathbf{z})$. We say the instance is *correctly assessed*, if y^* equals the actual state of disease of the instance, and *incorrectly assessed*, if the diagnosis was different. This accuracy measure could be easily improved by training a classifier on the features $\{X_j, Z_1, \cdots Z_k\}$ and predicting Y for each test instance.

The two approaches provide two different estimations of $I(X, Y|\mathbf{z})$. Our validation criteria (correctness and certainty) additionally requires the computation of $P(y|x_j, z_1, \ldots, z_k)$ and $H(Y|x_j, z_1, \ldots, z_k)$. The discussion in section 4.4 showed that it is not possible to compute this for large k. Therefore, we use the approximations shown in Table 4.2. H_P (H_Q) denotes the entropy of the probability distribution P (Q). To get a better grading of the results of our approaches, we compare it to two different ways of selecting a test.

We evaluate:

[3] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] We only consider instances with complete information on the test attributes as test instances (278 for heart disease and 138 for breast cancer) because it is not possible to evaluate the cases where the result value of the selected test is missing.

Ratio of correctly assessed instances						
	Heart Disease			Breast Cancer		
	prop.	*best*	*rand.*	*prop.*	*best*	*rand.*
Av.	0.781	0.996	0.681	0.657	0.846	0.696
J.M.	0.745	0.989	0.665	0.748	0.944	0.691

Table 4.3: Ratio of correctly assessed instances. The values in the columns differ in the way the test was chosen: proposed by information maximization, best possible (knowing the performance beforehand), or randomly selected. Note that *random* and *best* differ for each approach because of the different approximations of $P(y|x_j, \mathbf{z})$ (see. Table 4.2) used for the prediction.

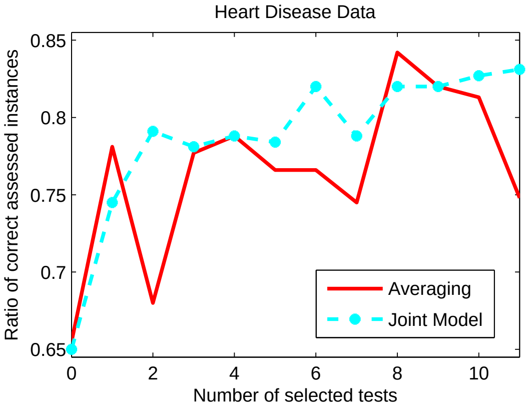


Figure 4.2: Ratio of correctly assessed instances when iterating the test selection. The graphs show the results of the two proposed approaches on the heart disease data.

- selecting a test *proposed* by information maximization
- selecting one of the possible tests at *random* (which is equal to averaging over all possible tests) against
- selecting the *best* test: we determine the performances of all tests beforehand and select the test with the best performance.

Because of the different approximations, the performance of the *random* and *best* methods differs for each approach.

Correctness Table 4.3 compares the ratio of instances that were assessed correctly: On both data sets holds, if we select an examination at random, in two thirds of the cases the decision about the status of disease is correct. On the heart disease data, selecting a test by information maximization leads to an improvement of about one tenth.

		$H(Y\mid x_j, \mathbf{z})$		
		Heart Disease		
		proposed	*best*	*random*
correct	Av.	0.733 ±0.15	0.613 ±0.14	0.862 ±0.07
	J.M.	0.615 ±0.24	0.418 ±0.21	0.801 ±0.13
incorrect	Av.	0.775 ±0.14	0.997 ±0.01	0.922 ±0.07
	J.M.	0.695 ±0.27	0.947 ±0.09	0.801 ±0.13
		Breast Cancer		
		proposed	*best*	*random*
correct	Av.	0.669 ±0.52	0.346 ±0.48	0.564 ±0.52
	J.M.	0.223 ±0.37	0.194 ±0.39	0.269 ±0.42
incorrect	Av.	0.731 ±0.58	1.258 ±0.14	0.591 ±0.59
	J.M.	0.499 ±0.43	0.786 ±0.51	0.599 ±0.44

Table 4.4: Mean value and standard deviation of entropy $H(Y|x_j, \mathbf{z})$. The lower the values the higher the **certainty** of the decision. The upper part of the table shows the average over correct assessed instances, the lower part over the incorrect assessed ones.

On the breast cancer data, selecting a modality by information maximization leads to an improvement in the second approach where we assume a simple joint model. Hence, the latter appears to give a more useful estimate of $I(X, Y|\mathbf{z})$. The difference of the two approaches on the heart disease data only shows when iterating the test selection process and adding the previous obtained test results as background knowledge. (Note: otherwise we have only two background attributes). The results are displayed in Figure 4.2. The second approach performs more stable and mostly outperforms the first approach.

Certainty of Decision To evaluate the certainty of the decision we evaluate the entropy $H(Y)$ of the distribution Y after a decision was made and an observation obtained. Table 4.4 shows the certainty of the decision split into the cases that were correctly classified (upper part of the table) and the incorrectly classified cases. On both data sets, we see that the certainty for the correct classified lesions is higher (lower entropy) than for the incorrect classified lesions. On the breast cancer data, the difference is large for the joint model approach, but for model averaging these quantities are very similar. This indicates that the proposed joint model is better suited at modeling the information content in the test/decision variables.

Actual Information Gain We compare the entropy of $P(Y|\mathbf{z})$ before performing a test with the entropy of $P(Y|x_j, \mathbf{z})$ after observing the test result $X_j = x_j$ (see Table 4.5). Comparing both approaches to random, the joint model approach achieves a better result than the baseline approach.

	Actual information gain							
	Heart Disease				Breast Cancer			
	Averaging		Joint Model		Averaging		Joint Model	
proposed	0.206	±0.14	0.275	±0.24	0.468	±0.53	0.549	±0.41
best	0.330	±0.15	0.460	±0.21	1.030	±0.31	0.660	±0.37
random	0.077	±0.05	0.109	±0.08	0.586	±0.29	0.327	±0.33

Table 4.5: Mean value and standard deviation of the actual information gain of observing test $X_j = x_j$: $H(Y|\mathbf{z}) - H(Y|x_j, \mathbf{z})$.

4.6 Related Work

In machine learning it is generally the case that learning can be made more efficient by appropriately choosing data points (from an available finite set); i.e., those data points that allow an algorithm to *learn* as much as possible [61]. The selection of new observations according to some criteria of interest can be thought of as form of *active learning*, also referred to as *optimal experiment design* or *sequential design* (*e.g.,* [26, 61, 59]). In the active learning literature multiple applications and numerous criteria have been proposed to address the problem of selection of new observations. However, active learning focuses on selecting entire instances that should be labeled and then added to the training set in order to improve the learned model. In this chapter, we are concerned about single feature values that should be observed for a new instance.

In contrast to active learning, *active feature-value selection* [64, 94] addresses the situation when having incomplete but labeled instances in the training set. The task is to select which instance should be completed to increase the performance of the model. Melville *et al.* [64] select instances that are missclassified by the current model and observe all missing values and relearn the model on the updated training set.

Gunter *et al.* [44] study how to select variables that actually influence the decision which action to take. Instead of ranking variables only due to predictive criteria, like in standard feature selection, they incorporate prescriptive criteria based on interaction of the variable with the action and the proportion of the population where the choice of decision depends on the variable.

Test selection has been investigated extensively in the medical and the statistical literature [7, 33, 40, 80, 93]. Doubilet [33] offers a mathematical approach to test selection, however, it assumes that prior probabilities can be computed or estimated, which is problematic for small training sets. Furthermore it is not clear how background knowledge can be incorporated. Andreassen [7] proposes to use causal probability networks (CPN) with utility nodes for test selection where the test is selected that increases the expected utility. However, this requires defining a CPN and utility functions. Glasziou and Hilden [40] discuss a variety of test selection measures and introduce some that incorporate costs. Sent and van der Gaag [80] investigate the behavior of several information measures for test selection and conclude that both, Shannon entropy and Gini index are well suited measures.

Our approach is based on the maximization of mutual information, a purely information theoretic criterion [83, 28]. In this context, several forms of experimental design have been studied dating back to [59] and more recently in [40, 61, 80], among others. Our approach is an application of these concepts. More specifically, we apply the concept of information maximization to select which random variable, among a finite set of variables, should be observed to maximize the expected information gain about another variable of interest. Since some variables are usually observed, we maximize the conditional mutual information instead.

4.7 Conclusion

We have employed the concept of mutual information to address the problem of choosing tests efficiently. We applied this to a problem in medical diagnosis. While mutual information is a well understood concept, in practice however, small datasets and underlying computational complexity make it difficult to calculate accurately for general probability models. We have experimentally shown how certain model assumptions can help circumventing these problems. Making these assumptions, we obtain a comparatively data-efficient variant of test selection based on information maximization. One limitation of this approach is that these models are build manually (not automatically), although model selection strategies and structure learning algorithms may be used. Our experimental results show that there is an advantage on conditioning on all the available observations even when conditional independence assumptions need to be made. Results indicate that the proposed joint model outperforms the information averaging approach by comparing the performance of each approach relative to a *random* and a *best* selection.

5 Subgroup Discovery for Test Selection

In this chapter, we propose a new approach to test selection based on the discovery of subgroups of patients sharing the same optimal test, and present its application to breast cancer diagnosis. Subgroups are defined in terms of background information about the patient. We automatically determine the best t subgroups a patient belongs to, and decide for the test proposed by their majority. We introduce the concept of prediction quality to measure how accurate the test outcome (in our case, the assessment of an imaging modality) is regarding the disease status (in our case, the biopsy). The quality of a subgroup is then the best mean prediction quality of its members (choosing the same test for all). Incorporating the quality computation in the search heuristic enables a significant reduction of the search space. We evaluate the approach on the data set of breast cancer diagnosis presented in Chapter 2.1. From a clinical point of view, our experiments show that the proposed approach provides valuable help at proposing optimal tests. From a computational point of view, our approach is faster than the baseline algorithm APRIORI-SD while preserving its accuracy.

5.1 Introduction

The goal of this work is to find the optimal set of tests to choose for a patient in a given situation, where the definition of optimality is provided in Section 5.2.1. Existing work on test selection [7, 33] mostly addresses the problem of finding global solutions for all patients. However, it is not likely that for each patient the same test is the most informative one. Therefore, we believe that it is a better approach to concentrate on the task of identifying subgroups of patients for which the optimal test is the same. Here, we present a novel solution to this problem based on subgroup discovery (SD) (see Chapter 3.2). Subgroup discovery methods compute all subgroups of a population that are statistically most interesting with respect to a specified property of interest. Standard SD approaches are designed for a single target variable. However, in the setting of test selection, a single variable seems not sufficient. In fact, we want to target the relation between two variables: the outcome of a test and the actual state of disease. Therefore, the quality of a subgroup should correspond to the value of the result of a selected test with respect to the actual state of disease, e.g., a biopsy result. To quantify the value of a test result, we define a so-called *prediction quality* function in Section 5.2.1. The function gives high scores to a pair of a subgroup and a test if the result is close to the actual state of disease, and therefore leads to an accurate diagnosis. Since standard SD does not take into account complex scenarios like this, including benefits or costs of subgroups, we developed a new, cost-sensitive variant.Throughout the thesis, we will use the term *prediction quality*, which corresponds to the *benefits* of a prediction rather than to its *costs*. However, as it is easy to transform one into the other, we can also speak of *cost-sensitive subgroup discovery*. The algorithm outputs subgroup descriptions consisting of background information about the patients. The overall goal is to compute an optimal test selection for a new patient. More precisely, our proposed solution is to identify subgroups of the data for which the same test is the optimal selection, to arrive at a correct diagnosis. In a second step, analyzing the subgroups will help to find out which features determine the performance of the tests. Hence, it will be possible to decide for a new patient, given its features, which test is the best to choose. We apply and validate this approach on a data set from breast cancer diagnosis, where for each patient four different tests are possible. The chapter is organized as follows: Section 5.2 recalls the setting of breast cancer diagnosis and the data set. The algorithm for cost-based SD is described in Section 5.3. The test selection procedure is evaluated in leave-one-out cross-validation. The results are described in Section 5.3.2. Section 5.4 discusses limitations of the data set. In Section 5.5, an overview of related work is given, and Section 5.6 concludes the chapter.

5.2 Background and Data

Our study was conducted in the area of breast cancer diagnosis. In breast cancer diagnosis, different imaging modalities are used routinely (see Section 2.1.1). Each modality has its own specific characteristics. When a patient is under scrutiny for breast cancer, it is often not clear which of these modalities is best suited to answer the basic question to whether the patient has or does not have cancer. The choice of a modality usually requires considerable experience of the health care workers. In this work, we show how to support the optimal test selection for a new patient by retrospectively analyzing the performance of the tests on subgroups of previously examined patients with similar features. The basis of our work is the dataset collected in a breast cancer study introduced in Section 2.1. It comprises patients that had a suspicious finding in a screening. The study gathers patient specific information like medical history, demographic information, and a breast cancer risk summary. Each patient in the study underwent all four above mentioned modality tests. Each of these tests was independently analyzed by the appropriate specialist to judge for the occurrence of breast cancer. For each lesion detected, the specialist determines in which category it falls. The categories are called BIRADS score and range from 0 to 5: The higher the BIRADS, the higher the probability (assessed by the medical expert) for the lesion to be malignant. (0 = incomplete, i.e., needs additional imaging evaluation, 1 = no finding, 2 = benign finding, 3 = probably benign, 4 = suspicious abnormality, 5 = highly suggestive of malignancy) [5]. To obtain evidence of the initial assessments, a biopsy has to be performed. A pathologic examination of a biopsy determines whether the lesion is benign, atypia benign, or malignant. In this study at least one lesion for each patient is subject to biopsy.

pscr		OA_m					
		0	1	2	3	4	5
BIO	Malignant	75	0	0	25	100	100
	Atypia	75	75	90	90	90	75
	Benign	75	100	100	100	75	50

Table 5.1: Prediction score for agreement between the overall assessment (OA_m = BIRADS) of a modality m and the biopsy finding BIO

5.2.1 Definition of Prediction Quality

To quantify the accuracy of a diagnosis, we propose a measure of prediction quality. Each test m results for each lesion l in an overall assessment $OA_m(l)$ of the physician. This is defined by the BIRADS score (see above). Each lesion has a biopsy $BIO(l)$ proving the status of the lesion (malignant, benign or atypia benign). The prediction quality expresses how close the assessment comes to the biopsy finding. Therefore, we define a prediction score *pscr* that evaluates the performance of a test for a single lesion. Table 5.1 gives the *pscr* for each pair (Overall Assessment, Biopsy). The higher the prediction score, the more accurate is the prediction. The values in the table were proposed by a domain expert in the field of breast cancer diagnosis. However, they are certainly not the only possible values for the prediction score. For instance, as not all types of malignant findings are equally harmful, it might be more accurate to distinguish between invasive and non-invasive types of cancer.

Having defined *pscr* for a single lesion l, we can easily obtain the prediction quality $pq(S, m)$ of a modality m for an example set S by averaging over the prediction scores of m and all lesions in S:

$$pq(S,m) = \frac{1}{|S|}\sum_{l \in S} \cdot\, pscr(OA_m(l), BIO(l))$$

test	DMAM	FMAM	MRI	USND
pq	77.9	78.0	78.4	80.2

Table 5.2: Prediction qualities of the modalities over all lesions

In our data set we have 138 lesions (of 72 patients) with biopsy and four modalities (Digital Mammography (DMAM), Film Mammography (FMAM), Magnet Resonance Imaging (MRI), and Ultrasound (USND)) to choose from. The prediction quality for the entire dataset separated for each modality is presented in Table 5.2. It shows that the prediction qualities of the different modalities over all lesions are quite similar (Entropy = 1.999 bits of a maximum 2 bits), with USND performing slightly better. By considering subgroups of patients we expect to increase the prediction quality for at least one modality per subgroup. Then, we apply this modality to all lesions in the subgroup to obtain the most accurate diagnosis.

5.3 Method

The general idea is to determine subgroups of lesions with an unusual modality performance. Let X be a training set of observed examples and n the number of tests $\{m_1, \ldots, m_n\}$ that can be performed. For each group[1] of lesions $S \subseteq X$ we consider the prediction qualities $pq(S, m_i)$ of the possible modalities and decide for the modality $m^*(S)$ with the highest pq-value[2]:

$$m^*(S) = \mathtt{arg}\max_m pq(S, m) \tag{5.1}$$

The optimal prediction quality of S is then defined as $pq^*(S) = \max_m pq(S, m)$.

We introduce an algorithm called $SD4TS$ (Subgroup Discovery for Test Selection). The task of the algorithm is defined in the following way:
Given: X, n, *minsupport* (the minimal number of examples that have to be covered by a subgroup description), t (the number of best subgroups we want to obtain from the algorithm), and a set of pq-values $\{pscr(s, m_i) | s \in X, m_i \in tests\}$ (in a more general setting a set of cost/benefit values).
Find: The t subgroups with the highest pq^* values (best costs/benefit) and at least *minsupport* examples.

We base our algorithm on APRIORI-SD [53], an adaptation of the association rule learning algorithm APRIORI [2] to subgroup discovery (see Chapter 3.2). APRIORI-SD starts with generating subgroups described by a single attribute-value-pair. Subsequently, it generates subgroups with longer (and thus more specific) descriptions. Subgroups are only kept if they contain more examples than *minsupport*. All smaller subgroups are pruned, and no subgroups more specific than these are generated. For our task, we are interested in the t subgroups that are cost-efficient for at least one modality. Therefore, we can prune the search space even further, namely in a way that only the promising subgroups are kept. That means, during the generation of subgroups, candidates are immediately evaluated and checked whether they have the potential to lead to improved costs.

5.3.1 Quality Pruning

Pruning is possible when a subgroup and all specializations of the subgroup will not outperform the quality of the already discovered subgroups. Specialization of a subgroup sg means adding an attribute-value pair to the subgroup description of sg. This can cause

[1] As in other SD algorithms we consider only groups that can be described by a conjunction of attribute-value pairs

[2] In a more general setting, instead of pq we can assume any types of costs (where max should be replaced by min) or benefits that rate the performance of the tests.

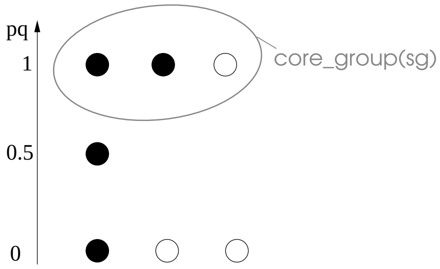


Figure 5.1: Simple pruning fails. All dots are examples in *sg*. The black dots are also covered by sg_{new}. The *y*-direction corresponds to the *pq*-value of each example.

changes of both the size and the quality of the subgroup. The size can never increase. The defined quality, however, can change in both directions.

The example in Figure 5.1 demonstrates the discussed characteristics. The seven dots represent the generated subgroup *sg*, with $pq(sg) = 0.5$. Assume we have generated already a subgroup sg_{best} with $pq(sg_{best}) = 0.6$. In this case, *sg* has a worse *pq* value and seems to be not promising. However, pruning *sg* will inhibit finding an optimal subgroup sg_{new} (the four black dots) contained in *sg* with $pq(sg_{new}) = 0.625$.

The critical point is hence to recognize when the quality of subgroup *sg* can not outperform at least one of the best *t* subgroups found so far. Thus, it is not enough to consider the actual $pq(sg)$ to determine if *sg* can be pruned. Furthermore, it is necessary to consider what we call the *coreGroup* of *sg*. The *coreGroup* is a group consisting of the *minsupport* examples covered by *sg* with the highest quality. The cost of the *coreGroup* upperbounds the costs of all possible specializations of *sg*, because the overall score is defined as an average of the elements of the group.

Considering the *pq*-value of the *coreGroup* of *sg* will circumvent this mistake by providing the upper bound of the *pq*-values of any specialization of *sg*: For the given example, we assume a *minsupport* of 3. Then $pq(coreGroup(sg)) = 1$. Since $pq(coreGroup(sg)) > pq(sg_{best})$, *sg* is not pruned and keeps the option of generating the improved subgroup sg_{new} in a later iteration of the algorithm.

5.3.2 The *SD4TS* algorithm

The pseudo-code of the algorithm is shown in Algorithm 1. *SD4TS* starts with generating 1-itemset candidates (described by one attribute-value-pair). Candidates are pruned if they are not frequent or if they can not outperform (not even through specialization) one of the the best *t* subgroups generated so far. All remaining candidates are stored in *optimizableCandidates*. The best *t* subgroups are stored in *topCandidates*. For efficiency, we store all created subgroups (including the list of transactions that are covered, costs, *bestPossibleCosts*, support and a list *still2test* of 1-itemsets that have not been tested yet to specialize the subgroup) in an array *allSubgroups* in the order they were created. The sorted lists *topCandidates* and *optimizableCandidates* contain only pointers (the indices of the array) to the subgroups stored in *allSubgroups*. The list *topCandidates* is sorted according to the actual costs of the subgroups. This facilitates removing the worst subgroup, whenever a newly generated subgroup has better costs. The list *optimizableCandidates* is sorted according to the *bestPossibleCosts* a *coreGroup* can achieve. In that way, we explore always the subgroup with the highest potential first. That means specializing this subgroup is likely to lead to a subgroup that falls into the top t candidates and therefore raises *minpq* which reduces the search space.

Algorithm 1 *SD4TS* (subgroup discovery for test selection)
Input: Training set, set of costs, *minsupport*, number t of best subgroups to be produced
Output: list of *topCandidates*, including the proposed test(s) for each candidate

```
$C$ ={$c$|$c$ frequent subgroup defined by 1 attribute-value-pair};
$topC$ = {$c$|$c$ belongs to the $t$ best candidates in $C$};
$minpq$ = worst quality of $topC$;
remove all $c$ from $C$ with $c.bestPossibleQ < minpq$
while $C$ not empty do
  $c_1 = C$.removeFirst();
  for all $c_2 \in c_1.still2test$ do
    $c_{new}$ = generate_new_candidates($c_1$, $c_2$);
    if $c_{new}$ frequent and ($c_{new}.bestPossibleQ > minpq$ or $|topC| < t$) then
      add $c_{new}$ to $C$
      if $c_{new}$ better than $minpq$ or $|topC| < t$ then
        add $c_{new}$ to $topC$
        if $size(topC) > t$ then
          remove worst $c$ of $topC$
        end if
        $minpq$ = worst quality of $topC$;
        remove all $c$ from $C$ with $c.bestPossibleQ < minpq$
      end if
    end if
  end for
end while
return $topC$
```

Safe pruning A new subgroup candidate sg_{new} is only accepted if sg_{new} is frequent and at least one of the following holds:

1. there are less than t subgroups stored in *topCandidates*, or
2. sg_{new} has a better performance than the worst subgroup of *topCandidates*, or

3. at least one frequent subset of examples in sg_{new} (i.e., $coreGroup(sg_{new})$) leads to a better performance than the worst subgroup of $topCandidates$. This can be tested in $O(n * |sg_{new}| \log |sg_{new}|)$: For each test m determine the set CG_m of $minsupport$ examples covered by sg_{new} that have the best costs. $sg_{new}.BestPossibleCosts = \max_m pq(CG_m, m)$

In all cases we add the candidate to *optimizableCandidates*. In case 1 and 2 we also add the candidate to $topCandidates$. In the second case we additionally remove the worst stored subgroup from $topCandidates$.

Eliminating duplicates Subgroup discovery algorithms often produce duplicate subgroups, that is, subgroups covering the same examples and differing only in the subgroup description. This arises mostly due to specializations of a subgroup s: Adding an attribute-value pair i_{new} to the description will create a new subgroup s_{new}, even if the covered examples are the same. $SD4TS$ eliminates duplicates. After each specialization it checks if the set of covered examples has changed (is smaller). If not, s_{new} will not be stored and only the description of s will be updated by flagging i_{new} as optional.

5.3.3 Different Subgroup Qualities

In this section, we introduce alternative scoring functions for subgroups that can be used in the pruning step of the algorithm. Let us compare two subgroups sg_1 and sg_2 of size n_1 resp. n_2, and the prediction qualities
$p_1 = (pq(sg_1, DMAM), pq(sg_1, FMAM), pq(sg_1, MRI), pq(sg_1, USND))$ and
$p_2 = (pq(sg_2, DMAM), pq(sg_2, FMAM), pq(sg_2, MRI), pq(sg_2, USND))$.
Then, possible meaningful scoring functions include:

1. Prediction quality of the best modality $pq(sg_1, BEST_1) > pq(sg_2, BEST_2)$
2. Relevance $= \sqrt{n_2} * \sqrt{\frac{n_1}{n_1 - n_2}}$;
   if $n_1 = n_2$: Relevance $= \sqrt{n_2}$
3. Information gain regarding choice of modality: $entropy(p_1) - entropy(p_2)$
4. New preferred modality, if $(arg \max_{p_1} \neq arg \max_{p_2})$;
5. General improvement of prediction quality: $\sum p_2 - \sum p_1$

In the implemented approach, we used the first criterion. However, for other scenarioes it might be useful to also test some of the other criteria, and modify the algorithm to prune according to other types of quality functions.

5.3.4 Analysis of Runtime and Search Space

Figure 5.2 shows how the search space (i.e., the number of search nodes) depends on the parameters *minsupport* and t. The higher *minsupport*, the smaller the search space. This is caused by the frequency pruning. We also see that a low t-value results in a small search space, which is the expected effect of quality pruning. For small values of t fewer subgroups are kept in *topCandidates*, which increases the threshold of costs below which subgroups are pruned. The right diagram in Figure 5.2 displays the runtime of the two algorithms. For *minsupport* values below 25, *SD4TS* is faster than APRIORI-SD, as frequency pruning is only effective for larger *minsupport* values.

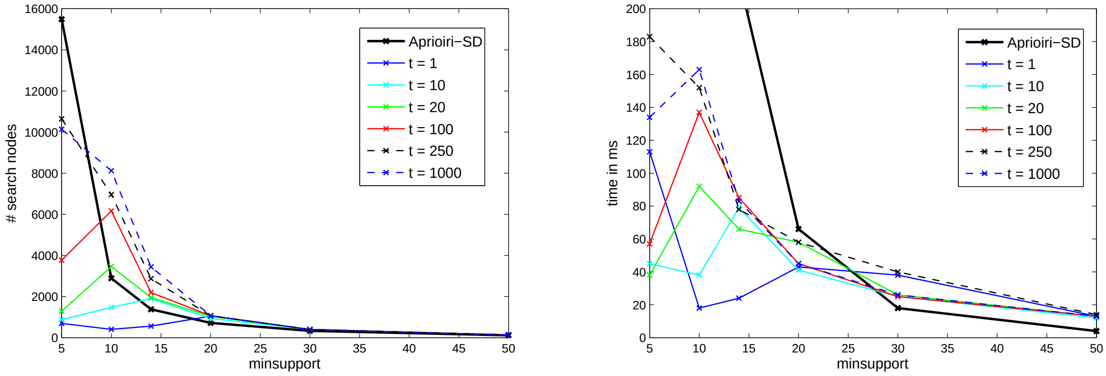


Figure 5.2: The left (right) diagram shows how the search space (runtime) depends on *minsupport* (x-axis) for different t-values. In comparison, the black solid line shows the search space of APRIORI-SD.

5.4 Validation and Results

To evaluate the approach, we tested it in a predictive setting[3], more specifically, in a leave-one-out cross-validation. For each test lesion l, we generate only subgroups with attribute-value pairs contained in l. Table 5.3 shows the best $t = 5$ subgroups for 3 example lesions.

From the resulting best t subgroups, we decide for the test proposed by the majority of the identified subgroups (for test lesion 9 it is USND). A test is proposed if it has the best costs averaged over all examples in the subgroup (for subgroup S1 it is USND). If more than one test has optimal costs, all of them are proposed (for subgroup S9 it is DMAM and USND). If more than one test is proposed most often by the subgroups, the cost for the test lesion l is determined by the mean of their costs.

[3] Note that prediction is not our main goal. Additionally, we are interested in the discovery of new medical knowledge. Therefore, we prefer subgroup discovery over standard classifiers.

	top 5 subgroups for selected test lesions	subgr. size	DMAM	FMAM	MRI	USND
	test lesion 9		0	0	0	**100**
S1	Highschool or less + has past Mammo	17	76.5	76.5	57.4	**98.5**
S2	Highschool or less + has past Mammo + no relatives with cancer	16	75.0	75.0	56.3	**98.4**
S3	Highschool or less + has past Mammo + has past breast USND	14	85.7	78.6	62.5	**98.2**
S4	Highschool or less + has past Mammo + ethnic group = white	14	85.7	78.6	66.1	**98.2**
S5	age 40-59 + no relatives with cancer + pre menopausal + ethnic group = white	28	76.8	72.3	76.8	**93.8**
	test lesion 19		75	100	100	**100**
S6	Graduate School after college + age 40-59 + no relatives with cancer + ethnic group=white	14	76.8	75.0	**98.2**	82.1
S7	Graduate School after college + has no breast USND + no relatives with cancer	21	94.1	82.1	92.9	**96.4**
S8	Graduate School after college + has no breast USND + no relatives with cancer + ethnic group = white	19	93.4	80.3	92.1	**96.1**
S9	Graduate School after college + age 40-59 + no relatives with cancer + has no breast USND	18	**95.8**	81.9	91.7	**95.8**
S10	Graduate School after college + has no breast USND+ no relatives with cancer + ethnic group = white + age 40-59	16	**95.3**	79.7	90.6	**95.3**
	test lesion 23		**100**	100	75	100
S11	no relatives with cancer + age $\geq$60	14	**94.6**	87.5	87.5	76.8
S12	Graduated from College + post menopausal status	16	78.1	82.8	**93.8**	70.3
S13	post menopausal status + age $\geq$ 60	15	**93.3**	86.7	86.7	76.7
S14	age $\geq$60	15	**93.3**	86.7	86.7	76.7
S15	Graduated form College + no relatives with cancer + post menopausal status	15	76.7	81.7	**93.3**	70.0

Table 5.3: Example of best 5 subgroups for 3 selected input test lesions. The shaded rows indicate the actual prediction scores of the modalities for the current input test lesion. The bold prediction qualities indicate the imaging modality proposed by *SD4TS*. For example, test lesion 9 will be assessed best by USND (pscr =100), the other three modalities fail to assess the lesion correctly (pscr = 0).

#best tests	# cases	code	# cases	specific code	# cases			
4 or 0	40 (29%)	1111	40	1111	40			
3	45 (33%)	1114	45	1114	10			Set 1
				1141	19			
				1411	7			
				4111	9			
2	21 (18%)	1133	16	1133	2			
				1313	4			
				1331	1			
				3131	1			
				3311	8		Set 2	
		1134	5	1134	1			
				1431	1			
				3411	1			
				4113	1			
				4311	1			
1	32 (23%)	1222	27	1222	2			
				2122	6			
				2212	11			
				2221	8	Set 3		
		1224	3	2124	2			
				2241	1			
		1234	2	2431	2			

Table 5.4: Distribution of lesions according to the ranking of modality performances. One modality outperforming three other equal modalities is encoded by 1222, etc. The first digit represents DMAM, the s second FMAM, the third MRI, and the last USND. The first column summarizes the codes according to how many modalities achieve the best performance.

5.4.1 Analysis of Performance Compared to Random Selection

For each lesion l, a vector of prediction qualities is given by

$$\overrightarrow{pq}(l) = (pq(l, FMAM), pq(l, DMAM), pq(l, MRI), pq(l, USND)).$$

This can be interpreted as a ranking of the modalities for each lesion. For instance, $\overrightarrow{pq}(l) = (0.8, 0.7, 0.9, 0.8)$ leads to the ranking $MRI > USND = DMAM > FMAM$. We encode this ranking as 1224. There is one modality ranked first, followed by two modalities ranked second, and one modality ranked fourth. In total, there are seven possible encodings: from 1111 (all modalities have the same prediction quality) to 1234 (all modalities have different prediction qualities). Additionally, we can encode in a more specific code which modality is best: The first digit refers to FMAM, the second to DMAM, the third to MRI and the last to USND.

costs	Set 1	Set 2	Set 3
SD4TS	**82.8**	**76.0**	**51.6**
Best	99.45	99.49	99.22
Random	**78.6**	**70.22**	**40.63**
Worst	58.03	41.33	16.41
DMAM	77.92	69.13	30.47
FMAM	78.1	69.39	41.41
MRI	78.28	69.9	44.53
USND	80.11	72.45	46.09

Table 5.5: Prediction qualities for different test selection methods. $SD4TS$ shows the results for the best parameter settings (see Table 5.6). Compare this to always picking the modality with the best costs, the worst costs, picking a modality at random, or always picking the same modality (always DMAM, always FMAM, always MRI, or always USND).

Table 5.4 shows the distribution of codes of our dataset. It is remarkable that in 29% of the cases all modalities perform equally well. This implies that for those cases a random choice is as effective as a more informed choice. To have fairer conditions, we additionally validated the algorithm on two restricted subsets of test lesions (results are shown in Table 5.6). Set 1 consists of all 138 lesions. Set 2 is a subset of Set 1 containing only the 98 lesions whose costs are not the same over all modalities (all codes except 1111). Set 3 comprises 32 lesions, where one modality outperforms the other modalities (code 1222, 1224, and 1234). Note that the differences between the best and the worst, and between the best and the random choice improve significantly from Set 1 to Set 3.

5.4.2 Results

We evaluated $SD4TS$ for different parameter settings. Results are shown in Table 5.6. It shows that the best parameter settings are low *minsupport* and high t values (considering even small subgroups), or, vice versa, high *minsupport* (50) and low t values (few large subgroups). Comparing the results in column *Set 1* of Table 5.6 with Table 5.5 shows that the algorithm achieves in general better costs than picking a modality at random or picking always the same modality. In general, the choice of parameters is good, if the prediction quality is higher than for a random selection. The results on the restricted sets *Set 2* and *Set 3* show that the improvement compared to a random selection increases from *Set 1* to *Set 3*.

We further validate the overall coverage of the lesions by the generated t subgroups. Figure 5.3a shows the percentage of lesions that are covered by at least one subgroup. With an increasing number of generated subgroups t, the coverage increases and the average quality (Figure 5.3b) decreases. It also shows that a higher minsupport induces a higher coverage, even for low values of t. Also for those cases the quality decreases. Figure 5.3c shows the behavior of the proportion of lesions covered by a subgroup when introducing a threshold, which has to be overcome by the prediction quality of the subgroups. Subgroups with lower prediction qualities are ignored in this setting. Of course with raising the threshold the number of uncovered lesions increases. Small *minsupport* allows more lesions to be covered. The average quality increases with a higher threshold for low *minsupport* and decreases for high *minsupport* and a higher threshold. The larger subgroups seem to be not specific enough.

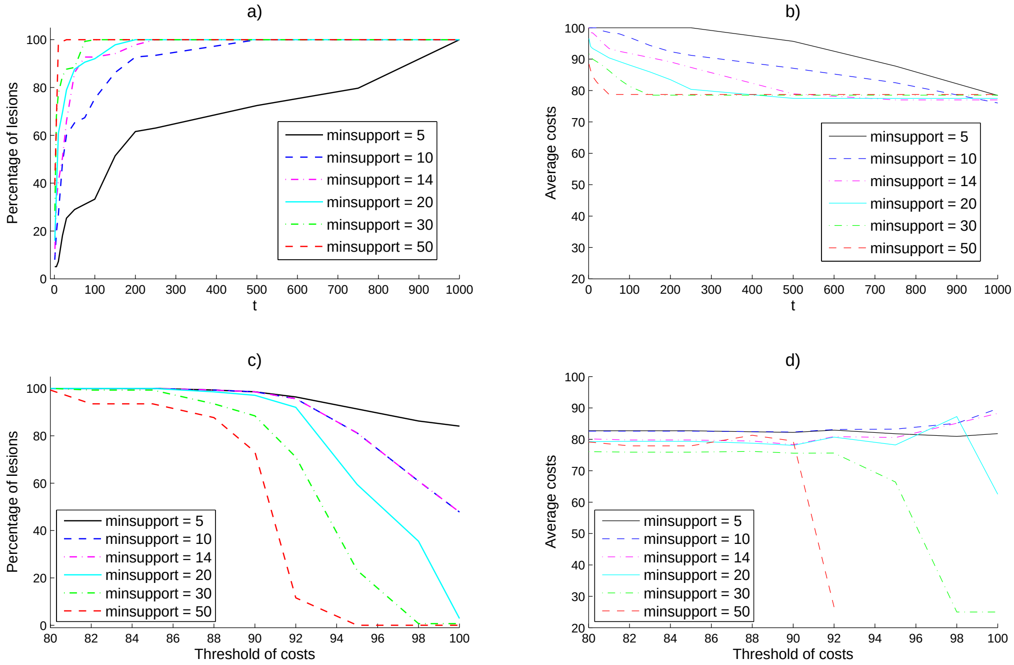


Figure 5.3: a) Percentage of lesions that are covered by at least one subgroup for different values of t (x-axis) and *minsupport*. b) Average costs (prediction quality) of lesions when choosing the modality proposed by the best subgroup ignoring lesions that are not covered by any subgroup. c) Percentage of lesions covered by at least one subgroup with pq above a certain threshold (x-axis). d) Average quality of lesions when choosing modality proposed by the majority of the (maximal 10) best subgroups above the threshold. Lesions covered by no subgroup above the threshold are ignored.

5.5 Related Work

Test selection has been investigated extensively in the medical and the statistical literature (see Section 3.1). However, none of the approaches allows identifying all subgroups with optimal costs. Other subgroup discovery algorithms [53, 9, 54, 56, 91] mostly focus on finding subgroups that are interesting or unusual with respect to a single target variable (mostly class membership; for numerical variables see [54]). In our problem setting we need a more complex target variable that expresses the relation between the test outcome and the biopsy. Exceptional model mining [58] provides an approach that is able to discover subgroups with a more complex target concept: a model and its fitting to a subgroup. It performs a level-wise beam search and explores the best t subgroups of each level. In contrast, $SD4TS$ does not require the definition of a model and is guaranteed to find the globally optimal subgroup. While subgroup discovery usually aims for a descriptive exploration of the entire population, we discover for each patient only subgroups that are supported by the patient's features. Therefore, we do not need a covering heuristic. With the introduction of prediction quality we have a measure that enables quality pruning of the search space (comparable to optimistic estimate pruning [91, 42]), whereas existing algorithms quite often only offer pruning according to frequency [53]. While test selection and subgroup discovery are well-investigated areas of research, their combination has not yet been considered in the literature.

5.6 Conclusion

Many questions in medical research are related to the discovery of statistically interesting subgroups of patients. However, subgroup discovery with a single target variable is rarely sufficient in practice. Rather, more complex variants, e.g., handling costs, are required. In this study, we considered such variants of subgroup discovery in the context of the clinical task of test selection and diagnosis: For our breast cancer diagnosis scenario, the task is to detect subgroups for which single modalities should be given priority over others, as indicated by a cost function. We designed an algorithm that handles costs and finds the most cost-efficient subgroups of a population. By limiting the output size to the best t subgroups, it is possible to prune the search space considerably, especially for lower values of the minimum frequency (i.e., support) parameter. Consequently, the proposed algorithm clearly outperforms the baseline algorithm used for comparison (APRIORI-SD) in our experiments. The main problems and limitations in the context of our study were caused by the small sample size (138 examples, i.e., lesions) and the non-unique solution for optimal test selection. In other words, for many cases, two or more tests perform equally well in practice. Our task could also be solved applying slightly modified methods from Section 5.5. However, we expect this to result in longer runtimes. Overall, we showed that subgroup discovery can be adapted for test selection. Similar techniques should be applicable successfully in other areas as well. In the next chapter, we compare the method with the information-theoretic approach of test selection discussed in Chapter 4.

Set 1						
	min-s.	5	10	20	30	50
t	1	80.7	**82.6**	77.8	79.2	**81.5**
	5	79.5	81.3	76.6	78.2	**82.7**
	10	**82.7**	80.7	76.1	79.0	80.1
	20	**82.6**	79.6	77.8	80.4	80.5
	30	**82.5**	79.3	80.1	79.2	80.5
	100	**82.8**	80.3	79.7	79.1	80.5
	250	**82.2**	80.7	79.7	79.1	80.5

Set 2						
	min-s.	5	10	20	30	50
t	1	73.1	**75.8**	69.1	70.9	**74.2**
	5	71.4	73.9	67.4	70.2	**75.9**
	10	**75.9**	73.1	66.6	70.7	72.2
	20	**75.8**	71.5	69.0	72.7	72.8
	30	**75.6**	71.1	72.2	70.9	72.8
	100	**76.0**	72.6	71.7	70.8	72.8
	250	**75.2**	73.1	71.7	70.8	72.8

Set 3						
	min-s.	5	10	20	30	50
t	1	45.6	48.4	37.9	42.6	47.3
	5	42.8	42.1	35.6	43.0	**51.6**
	10	47.3	41.8	37.1	46.9	42.5
	20	44.5	42.6	42.6	49.6	47.1
	30	49.6	46.1	49.2	41.4	47.1
	100	**51.0**	47.7	46.1	41.4	47.1
	250	**51.0**	47.7	46.1	41.4	47.1

Table 5.6: Results of leave-one-out cross-validation of *SD4TS* with varying minsupport and *t* parameters over three different test sets (best parameter settings in **bold**). The displayed costs are averaged over all test lesions. A cost for a single lesion is derived by taking the costs of the test proposed by the majority of the returned subgroups. If two or more tests are proposed equally often, we take the average of their costs.

6 Comparison of Test Selection Methods

In the previous chapters, we introduced two approaches to test selection. We refer to them as $IMAX$ and $SD4TS$. In this chapter we compare the two methods. For the experiments in Chapter 4 and Chapter 5 we considered only the 138 lesions with complete information on the test attributes and the biopsy findings. Therefore, we restrict our analysis in this chapter to these 138 lesions.

6.1 Which Modalities are Selected?

First, we compare which modalities are proposed by the two approaches for the 138 lesions. Figure 6.1 shows the distribution of modalities selected by $IMAX$ and by $SD4TS$. We see that $IMAX$ prefers MRI whereas $SD4TS$ prefers USND.

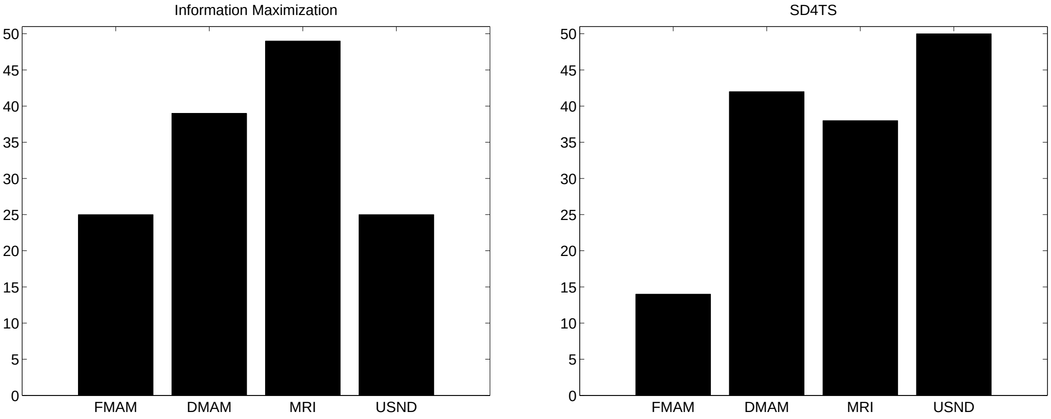


Figure 6.1: Distribution of selected modalities for $IMAX$ (left) and $SD4TS$ (right).

6.2 How Accurate is the Overall Assessment compared to the Biopsy?

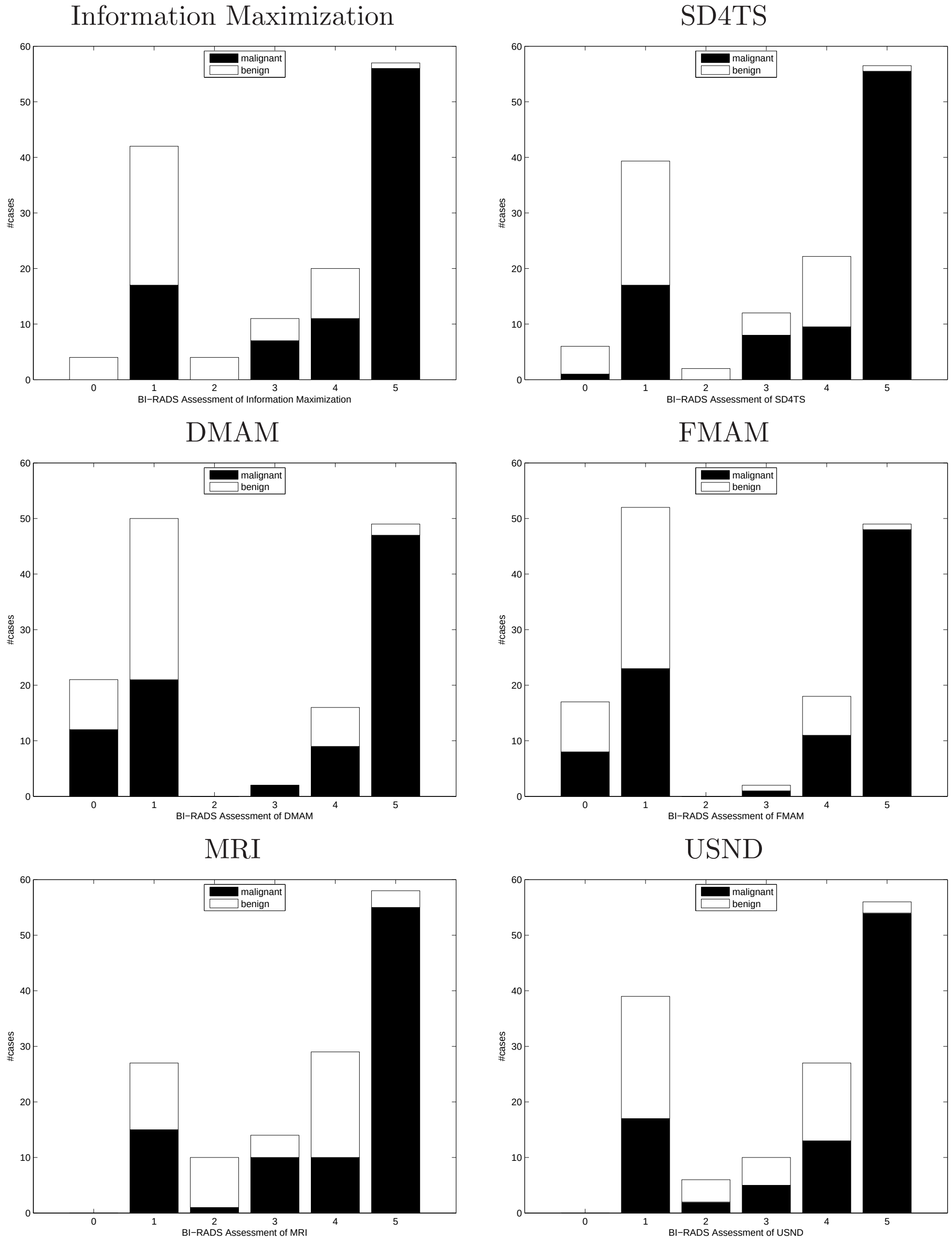
Next, we compare the accuracy of the diagnosis obtained by selecting the proposed image modality. As in Section 2.1.2, we illustrate the accuracy of the overall assessment compared to the biopsy findings. Table 6.1 displays the results for the modalities selected by $IMAX$, $SD4TS$, and for the case of always selecting the same modality. It shows that $IMAX$ and $SD4TS$ do similarly well with some variations in BI-RADS=2 and BI-RADS=4. Both x-ray methods ($DMAM$ and $FMAM$) do worse, i.e., they evaluate more malignant lesions evaluated as benign (with BI-RADS=1). Compared to MRI and $USND$, our proposed methods evaluate fewer benign lesions a malignant (BI-RADS=5).

We also compare the prediction quality *pq* of both approaches: $SD4TS$ yields a prediction quality of 82.8 and $IMAX$ a slightly lower one of 81.6. Comparing this to the results for $Set1$ in Table 5.5 and Table 5.6 shows that $IMAX$ can compete with the best parameter settings for $SD4TS$ and outperforms picking always the same modality or picking a modality at random.

On the other hand, we validate $SD4TS$ by the *correctness* criterion from Section 4.5, which is the ratio of correctly assessed lesions, when applying the selected test and deciding for the most probable biopsy outcome given the test result and all available background attributes. We determine the probability by the probability distribution Q of the approach that assumes a joint model. This yield a correctness of 0.733 for $SD4TS$ which is slightly lower than the correctness for $IMAX$ of 0.748 but still much higher than selecting a modality at random (compare Table 4.3).

6.3 Conclusion

We could show that both of our approaches perform better than selecting a methodology at random. The two approaches achieve similarly good results. The main difference is that $SD4TS$ also provides a descriptions of subgroups that have the same optimal test, which makes it possible to explain relations of the patients features and the selected modality and hence discover novel medical knowledge. In contrast, $IMAX$ is especially adapted for situations when physicians plan to apply a chain of modalities. For instance, when the outcome of a test is already known, $IMAX$ is able to incorporate this knowledge as a background attribute. This enables a better decision for an additional image methodology.



Information Maximization

SD4TS

DMAM

FMAM

MRI

USND

Table 6.1: Accuracy of BI-RADS-Assessments (X-axis) of single modalities (DMAM, FMAM, MRI and USND). The Y-axis gives the number of cases of a certain x-value. The color of the piles encodes the portion of lesions with the same biopsy result.

7 Subgroup Discovery on Clusterings

In the following, we present a new approach to subgroup discovery with more complex output in the form of images. The study was conducted in the medical domain of Alzheimer's disease and other forms of dementia. The psychiatry and nuclear medicine departments of Klinikum rechts der Isar (TU München) granted access to psychological and physiological data of patients suffering from Alzheimer's disease or dementia. The data consists of PET scan images and structured data containing patient information and the results of psychological tests. A detailed description can be found in Section 2.2.

One of the goals of medical research in the area of dementia is to correlate images of the brain with clinical tests. Our approach is to start with the images and explain the differences and commonalities in terms of the other variables. First, we cluster PET scans of patients to form groups sharing similar features in brain metabolism. To the best of our knowledge, it is the first time ever that clustering is applied to whole PET scans. Secondly, we explain the clusters by relating them to non-image variables. To do so, we employ an algorithm for relational subgroup discovery, RSD, with the cluster membership of patients as target variable. Our results enable interesting interpretations of differences in brain metabolism in terms of demographic and clinical variables. The approach was implemented and tested on an exceptionally large data collection of patients with different types of dementia. It comprises 10 GB of image data from 454 PET scans, and 42 variables from psychological and demographical data organized in 11 relations of a relational database. We believe that explaining medical images in terms of other variables (patient records, demographic information, etc.) is a challenging new and rewarding area for data mining research.

7.1 Motivation

Given the results from neuroimaging studies, one of the major goals of medical research is to correlate them with other non-image based variables (e.g., demographic information or clinical data). The usual approach is to select a subset of patients fulfilling specific predefined criteria (e.g., the level of cognitive impairment) and to compare the images associated with those patients to data from healthy controls in a group analysis. However, it is clear that such an approach can never be guaranteed to be complete: If the first step misses an important subset, it is not possible to recover from this omission in subsequent steps.

Therefore, we propose to apply data mining techniques to take the opposite approach: to start with the images and explain the differences and commonalities in terms of non-image variables. In this way, the results of the analysis are less dependent on the choices made in the selection of patients. In fact, the goal is to obtain a complete list of descriptions of subgroups of patients, which are unusual with respect to the PET images. The approach was implemented and tested on data derived from an exceptionally large pre-existing data collection of patients with different types of dementia, collected at our university hospital. In the first step, we clustered PET scans of patients to form groups sharing similar features in brain metabolism. In the second step, we explained the clusters by relating them to clinical and other non-image variables. To do so, we employed RSD [57, 86], an algorithm for relational subgroup discovery, with the cluster membership of patients as the target variable. After extracting relevant information from 200 GB of data (removing duplicates, intermediate results, and incompletely processed images), we obtained a dataset comprising 10 GB of image data from 454 PETs, and 42 variables from clinical and demographical data organized in 11 relations of a relational database. Large image clusters identified metabolic patterns corresponding well to typical findings in major types of dementia. Furthermore, the approach allowed the detection of differences in cognitive performance in presence of comparable brain pathology, thus potentially helping to identify factors supporting compensation (e.g., age, gender, education).

In summary, the contributions of this chapter are as follows: First, we present a new application area and task for data mining in a highly relevant area of medical research. Second, we present the first clustering of whole PET scans. Third, we propose, motivated by medical considerations, a new type of correlation analysis based on a loose coupling of clustering and subgroup discovery. Fourth, the procedure itself is novel in the medical area, as the approach is diametral to current practice in the analysis of PET images and deemed highly relevant by medical experts in the field.

This chapter is organized as follows: In Section 7.2 we explain our workflow and how we applied the clustering and subgroup discovery algorithms to the data set. Section 7.3 presents our results and their interpretation by medical experts. Section 7.5 discusses the results from a higher perspective. The chapter closes with a review of related work (Section 7.6) and the overall conclusion.

7.2 Method

The data mining part of our approach is illustrated on the right-hand side of Figure 7.1. In the first step, we apply k-Medoids clustering to the image data. The clusters of the best clusterings according to clustering quality and expert evaluation are further interpreted by the corresponding non-image data. For the interpretation of image clusters, we employ RSD (relational subgroup discovery) [57, 86], an algorithm for finding interesting subgroups in data. In subgroup discovery, the goal is to find subgroup descriptions (typically conjunctions of attribute values as in rule learning) for which the distribution of examples with respect to a specified target variable is “unusual” compared to the overall target distribution. In our case, clinical and demographic variables form the subgroup descriptions, and the cluster membership is chosen as the target variable. Before applying RSD, we still remove images with incomplete corresponding non-image

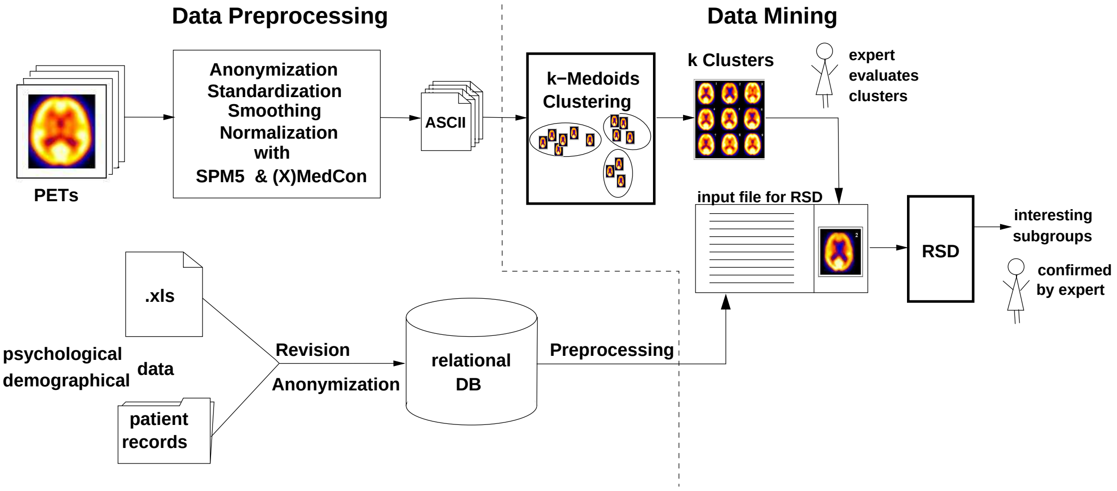


Figure 7.1: Workflow of our approach. The upper part shows the processing of image data, the lower part the processing of the structured non-image data. Preprocessing steps are on the left side, data mining and interpretation of results on the right.

data as well as all clusters below a certain size.

In the following, we give a short description of the clustering and subgroup discovery algorithms and explain how we applied them to our data.

7.2.1 Clustering

We chose the k-Medoids algorithm [52], a derivate of k-Means [92], with variance-weighted features to cluster PET scans. Alternative clustering approaches and their results will be discussed in Section 7.5. As mentioned above, the images have to be normalized and standardized before clustering. The resulting ASCII files (with over 500,000 real-valued entries) were taken as input for k-Medoids. To obtain meaningful and significant clusters using k-Medoids, it is necessary to weight the features according to their variance. This is feasible due to the huge differences in the variance of the intensity in different brain regions. These differences are caused, among others, by activation patterns specific for certain types of dementia and for healthy controls. Moreover, it is clear that not all parts of a PET scan reflect the state of brain tissue: For instance, in the PET image, the brain is surrounded by a dark area which is the same for all scans (cf. Figure 2.8). Figure 7.2 shows the variance distribution of the voxels in the 32nd layer over all PET scans. As regions of low variance are not very informative for the clustering, we chose to weight each voxel at position (i, j, k) by the standard deviation over all PET scans s_{ijk}. Therefore, we obtain the following distance measure based on

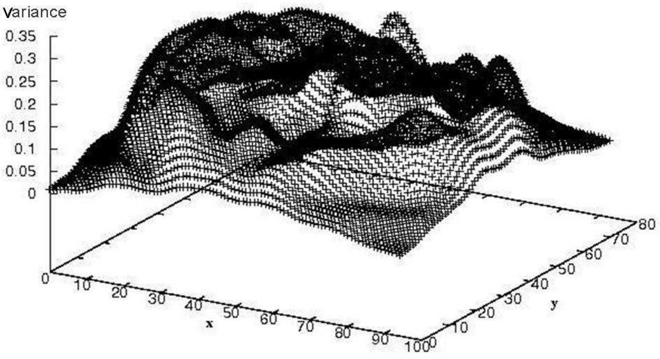


Figure 7.2: Distribution of variance (over all patients) of voxels from the 32nd layer of the PET scans.

the Euclidian distance measure with the variables weighted by their standard deviation:

$$dist(A, B) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} (s_{ijk} \cdot a_{ijk} - s_{ijk} \cdot b_{ijk})^2} \tag{7.1}$$

Since the optimal number k of clusters is not known in advance, we need a quality measure to compare clusterings with different k. One of the few quality measures for clusterings independent of k is the silhouette coefficient (SC) [52]. For a single cluster C of a clustering $\mathcal{C}$ it is defined as:

$$SC_C(\mathcal{C}) = \frac{1}{|C|} \sum_{x \in C} \frac{b(x, \mathcal{C}) - a(x, \mathcal{C})}{max\,\{a(x, \mathcal{C}), b(x, \mathcal{C})\}} \tag{7.2}$$

where $a(x, \mathcal{C})$ is the distance of object x to the medoid of the cluster C it belongs to and $b(x, \mathcal{C})$ the distance of object x to second nearest medoid. For an entire clustering $\mathcal{C}_k$ we used the mean of $SC_C(\mathcal{C}_k)$ over all C which is referred to as $SC(\mathcal{C}_k)$ in the following.

It always holds $-1 \leq SC \leq 1$. A high SC does not necessarily reflect the best clustering, since $SC(\mathcal{C}_k) = 1$ for all $\mathcal{C}_k$ with $k = 1$ or $k = |X|$. Generally, $SC(\mathcal{C}_k)$ increases with $k \to |X|$, because $SC_C(\mathcal{C}_k) = 1$ for all clusters C that consist of one example only.

On our data, clustering with more than 30 clusters results in too small clusters despite their high SCs. Clustering with $k < 10$ is not informative as well, because it tends to find results consisting of two or three very large clusters and single outliers forming their own clusters. In this case, the resulting $SC(\mathcal{C})$ is high because outliers forming their own clusters have an SC of 1, which increases the overall SC. Therefore, the number of appropriate k was expected to be between 10 and 30.

As the results of k-Medoids depend both on the initial choice of medoids and the input order of objects, it was tested 5000 times for each $k \in 2, \ldots, 100$, and the clustering $\mathcal{C}_k^*$ with the maximum SC was chosen. The initial computation of the distance matrix needed two hours, and the 5000 runs took two hours for all k on a 1GB RAM (1.6 GHz) machine.

For the obtained SC distribution, local maxima exist at $k = 10$, 12, and 16. We hence chose to further examine these three clusterings using subgroup discovery on the corresponding clinical and demographic data. Since our analysis showed very similar results for clustering with $k = 12$ and $k = 10$, we omit the presentation of the results for $k = 12$, and focus on only those for $k = 10$ and $k = 16$ in Section 7.3.

To determine the significance of a clustering with parameter k, an additional randomization test was performed. All original scans were shuffled to obtain 454 new images in a way that each voxel occurred again in one of the new scans, while keeping its location. So, the variance for each voxel stays the same. Subsequently, a clustering for parameter k was performed 100 times similarly to the original clustering. The resulting SCs were used to create a sorted list. For a given clustering $\mathcal{C}_k$ with original data, the p-value can easily be determined by the rank of the corresponding silhouette coefficient $SC(\mathcal{C}_k)$ in this list. The final validation of the clustering is done by presenting the mean images to an expert, who interpreted them and explained details.

7.2.2 Subgroup Discovery

Subgroup Discovery is a method for finding subgroups in a dataset that are sufficiently large and statistically unusual given their distribution on an attribute of interest (see Chapter 3.2 for an introduction). In this work, we used the subgroup discovery algorithm RSD (see Section 3.2.1) and its publicly available Prolog implementation[1] [85].

The parameters of RSD were set as follows: For getting the maximal number of interesting subgroups for each class, the output was increased to 20 subgroups for each class. The beam width was set to 15, and the maximal number of literals in each rule was set to 4. For the clustering results with $k = 16$, the running time of RSD on a Pentium 4 (2,8 GHz) with 1 GB RAM was approximately 8 hours.

The significance of the subgroups was determined by the likelihood ratio (see Equation 3.4), where k is the number of clusters. Interesting subgroups were identified by checking their p-value[2] and by expert validation.

7.3 Results

This section focuses on the clusterings $\mathcal{C}^*_{10}$ and $\mathcal{C}^*_{16}$. First, we present the medical interpretation of the mean images of the obtained clusters in Section 7.3.1. Subsequently, we discuss in Sections 7.3.2, Section 7.3.3 and 7.3.4, the characteristics of clinical values and subgroup results from RSD with the clusters from the clusterings with $k = 10$ and $k = 16$ as the target variable. Moreover, we relate the subgroup descriptions to medical expert knowledge.

[1] http://labe.felk.cvut.cz/~zelezny/rsd/

[2] A subgroup is considered significant, if its p-value is below 0.05.

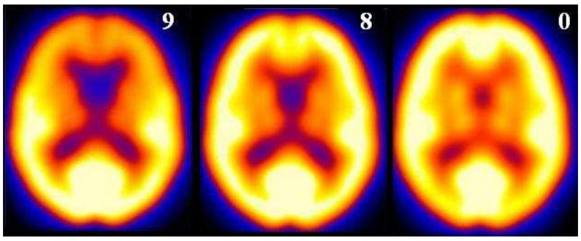


Figure 7.3: Mean images of the three largest clusters of the clustering with $k = 10$.

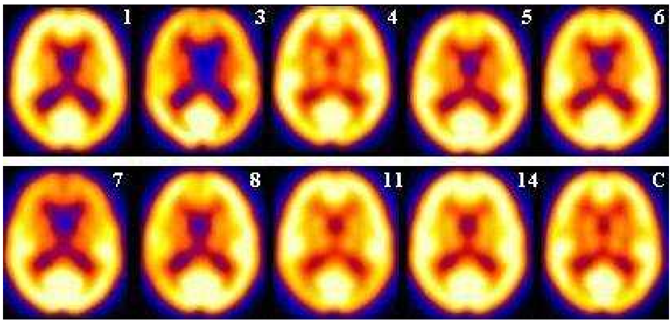


Figure 7.4: Mean images of the nine largest clusters of the clustering with $k = 16$. In the second row on the right is the mean image of the healthy control group C.

7.3.1 Clustering of PET Scans

Both clusterings (with $k = 10$ and $k = 16$) were significant with a p-value less than 0.01 as determined by the randomization test. Generally, the clusters vary widely in their size and also encompass singleton clusters, i.e., outliers. More specifically, the distribution of cluster sizes for clustering $\mathcal{C}^*_{10}$ is (187, 2, 5, 2, 4, 1, 5, 1, 207, 40), meaning that the first cluster consists of 187 PETs, the second of 2, and so forth. In the following, we refer to the first cluster in this list as Cluster 0, and the i-th as Cluster $i - 1$. Analogously, the distribution of cluster sizes for clustering $\mathcal{C}^*_{16}$ is (2, 42, 1, 8, 105, 104, 28, 40, 8, 1, 3, 61, 1, 1 , 48, 1). Thus, both clusterings found outliers and other very small clusters. In fact, one image that was sorted into its own cluster was rotated upside down. It was therefore most different from all other scans. The other singletons scans were strongly deformed, for which SPM5 is unable to compensate. For instance, some patients kept their chin too close to the chest during the recording.

For all clusters containing at least five PETs, we calculated their mean images (five for clustering with k=10, and nine for clustering with k=16). In both clusterings we found clusters grouping patients with frontotemporal dementia (Cluster 9 in Figure 7.3 and Cluster 7 in Figure 7.4), nearly healthy patients (Cluster 0 in Figure 7.3 and Cluster 4 in Figure 7.4) and global hypometabolism (Cluster 3 in Figure 7.4 and Cluster 2 of $\mathcal{C}^*_{10}$ not illustrated). Clustering with k=16 performed better, because it managed to differentiate more precisely the group of left lateral deficiency (Cluster 8 in Figure 7.4) and a typical

Alzheimer's cluster (Cluster 11 of size 61 in Figure 7.4). Both were not separated by the clustering with k=10. Even though the remaining clusters cannot be interpreted clearly from a medical point of view, they can be distinguished visually. This can be seen by the differences in metabolism in the occipital and centering regions, which are lighter in Cluster 5 than in Cluster 6 (Figure 7.4). Cluster 1 shows highly affected patients.

In summary, the clusterings were judged as meaningful by domain experts (R. Perneczky and A. Drzezga). To further explain the differences in the cognitive areas, we combined the images with clinical data. In the next section, we present a simple correlation analysis and relate the clusters to single non-image variables. This serves as a baseline for the more complex approach based on subgroup discovery.

7.3.2 Simple Correlation Analysis of $\mathcal{C}^*_{16}$

In medicine, next to the diagnosis and test results, age and gender are important variables to describe status and progression of a disease. Therefore, we first relate the large clusters of $\mathcal{C}^*_{16}$ to those attributes. Looking at the distributions of diagnoses of $\mathcal{C}^*_{16}$, some clusters have a high proportion of patients with Alzheimer's disease, while others contain only patients with a cognitive disorder (Figure 7.5 clusters 4 and 11). This indicates that k-Medoids clustering is capable of grouping similar images together. This is also supported by the MMSE values of the clusters (Figure 7.6). Cluster 4 has the highest MMSE score and a low variance, indicating that its patients are almost healthy. In contrast, the MMSE of Cluster 14 is not bad (around 25), but the high variance indicates that there exist some patients that suffer from a more severe disease. The same is true for clusters 8, 11 and 1. This is an interesting finding, because it states that patients with a similar brain metabolism may have different cognitive abilities. Cluster 8 comprises patients which have a left lateral metabolic deficiency, so the low MMSE score can be explained well. Again, the high variation states that some people might not be impaired too much from this hypometabolism.

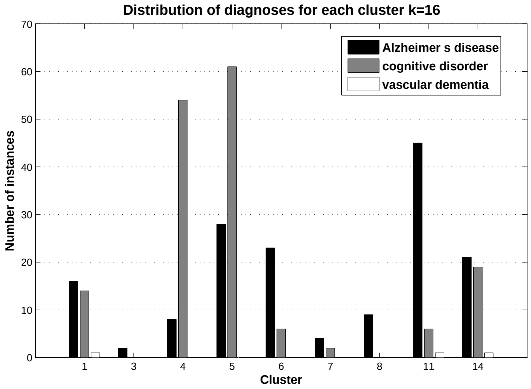


Figure 7.5: Distribution of ICD10 codes for clusters of $\mathcal{C}^*_{16}$

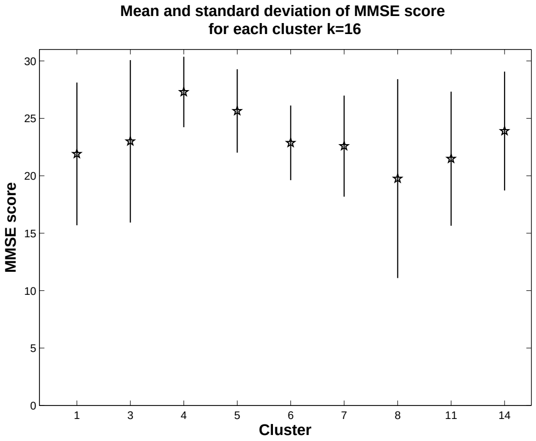


Figure 7.6: Distribution of the MMSE score for clusters of $\mathcal{C}^*_{16}$

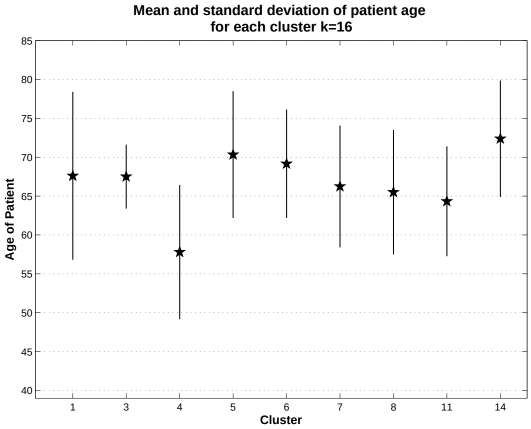


Figure 7.7: Distribution of age for clusters of $\mathcal{C}^*_{16}$

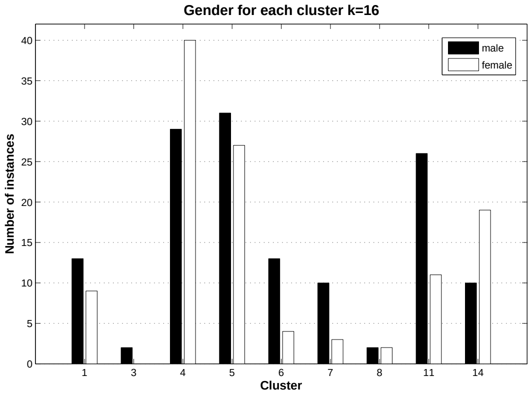


Figure 7.8: Distribution of males/females for clusters of $\mathcal{C}^*_{16}$

As age is correlated with the progress of dementia, we also investigated the distribution of age among the clusters (Figure 7.7). The difference between Cluster 14 and Cluster 4 is approximately 20 years. The average age of Cluster 14 is around 75 while Cluster 4 has younger people of age 55, which goes hand in hand with the distribution of ICD10 codes that define Cluster 4 as an almost healthy cluster. Cluster 3 also has a very high average age and a low variance characterizing old morbid people with an advanced hypometabolism.

Concerning the gender distribution of each cluster, Cluster 4 and 14 have a higher fraction of women, while Cluster 11 holds more men (Figure 7.8). This is quite interesting, because it suggests that there might be some differences in the development of dementia between the genders. Another explanation may be a different behavior as to when a physician is consulted.

This first overview of attributes in the clusters shows their general characteristics. However, simple correlation analysis is not able to detect dependencies among the attributes. Clearly, the correlation of multiple attributes can only be analyzed by more complex methods like subgroup discovery. It allows us to explore the interaction of attributes within a cluster. For instance, Cluster 14 has a high variance in MMSE scores and an unbalanced gender distribution, but so far we cannot see if men have a better MMSE score than women, or if there is no significant difference between the genders. Identifying subgroups helps to combine interesting characteristics and allows to take into account several variables at once.

7.3.3 Subgroup Discovery on $\mathcal{C}^*_{10}$

Our application of subgroup mining requires reliable psychological data and sufficiently large clusters. Thus, we first discard those examples without revised psychological data. Next, we eliminate all clusters with less than 5 examples after the first filtering step. Thus, we keep Cluster 0, Cluster 8, and Cluster 9, with the sample distribution $(115, 118, 14)$ (initially, it was (187, 207, 40)). Figure 7.3 shows the mean images for each of these clusters.

Table 7.1 displays the interesting subgroups discovered for Cluster 0. Each row in the table presents a subgroup description along with quality measures and the distribution of cases over the clusters. For instance, the first subgroup A1 covers 29 images from Cluster 0, 5 images from Cluster 8, and none from Cluster 9.

The mean image of Cluster 0 resembles the 20 healthy controls (Figure 2.8), which leads to the assumption that this cluster contains mainly almost healthy patients. Subgroup A1 shows that 85% of the patients, younger than 55, fall into Cluster 0. Experts confirm that the younger a patient, the better the activity of metabolism and the lower the probability of a dementia diagnosis. Furthermore, subgroup A2 and A9 reveal that the distribution of patients with good test results is especially high in Cluster 0, which also explains the similarity of the mean image to the healthy controls. Hence, these subgroups confirm the assumption that Cluster 0 contains the healthier patients.

However, there is also one subgroup of patients with Alzheimer's diagnosis (A4) and one describing patients with a very low score for *CDR activities* and *CERAD construc-*

Id.	**Description**	**p-value**	$\frac{\lvert sg \cap cl \rvert}{\lvert cl \rvert}$	$\frac{\lvert sg \cap cl \rvert}{\lvert sg \rvert}$	**Class distr.**
A1	age $\leq$ 54	$< 10^{-4}$	0.25	0.85	(**29**,5,0)
A2	verbal fluency $\geq$ 20	7.10^{-3}	0.22	0.64	(**25**,11,0)
A4	ICD = Alzheimer's & age: 65-69	5.10^{-2}	0.13	0.7	(**14**,6,0)
A9	MMSE: 26-29 & BNT: 14-15 & ICD = F06.7 & age: 55-59	2.10^{-2}	0.05	1	(**6**,0,0)
A11	activities = 1 & constructional praxis: 3-5	5.10^{-2}	0.04	1	(**4**,0,0)
overall distribution over clusters **0**, 8, 9:					(**115**,118,14)

Table 7.1: Interesting subgroups in Cluster 0 of $\mathcal{C}^*_{10}$. $\frac{|sg \cap cl|}{|sg|}$ is the proportion of patients in the subgroup sg having a PET in cluster cl. In contrast, $\frac{|sg \cap cl|}{|cl|}$ expresses the proportion of the patients with a PET in cluster cl that are also covered by the subgroup description of sg.

tional praxis (A11). Further research revealed that these people suffer from Alzheimer's disease in an advanced state, which should not fall into this cluster. The corresponding mean images of the outliers in this cluster had the most affinity to Cluster 0 in high-variance areas. With the choice of a larger k, the algorithm can sort them out into different clusters. We can conclude that the subgroup mining step identifies four outliers within Cluster 0.

Cluster 9 (in Figure 7.3) describes patients that have a huge deficit in the frontotemporal metabolism. This leads to the conclusion that they are not affected by Alzheimer's disease, but by some other form of dementia. Furthermore, experts assumed no significant reduction of *CERAD constructional praxis* (A41), while *CERAD verbal fluency* is highly impaired (A55) as well as scores in *CDR* (A42). These conclusions could be confirmed by the subgroups in Table 7.2.

The discovered subgroups for Cluster 8 showed that it consists mainly of patients above the age of 74. However, for a medical interpretation of this cluster, the medical experts could not obtain additional information from the mean image or the other subgroup descriptions. Thus, we assume that this cluster is too heterogeneous and that we need a clustering with a larger k to split this cluster into smaller, more homogeneous clusters.

Id.	Description	p-value	$\frac{\lvert sg \cap cl\rvert}{\lvert cl\rvert}$	$\frac{\lvert sg \cap cl\rvert}{\lvert sg\rvert}$	Class distr.
A41	constructional praxis: 10-11 & gender: m & ICD: other diag.	4.10^{-3}	0.57	0.22	(14,14,**8**)
A42	memory: 1 & activities: 1 & judgment: 0-0.5 & community: 1	2.10^{-3}	0.29	0.5	(3,1,**4**)
A49	community: 2-3	2.10^{-3}	0.29	0.4	(1,5,**4**)
A52	CDT: ≤ 2 & word list recall: 10 & verbal fluency: 9-12 & other tests: CT	2.10^{-4}	0.21	1	(0,0,**3**)
A53	CDT: 3-6 & global: 2-3 & community: 2-3	2.10^{-3}	0.21	0.6	(0,2,**3**)
A55	gender: m & verbal fluency: 0-8 & community: 2-3	2.10^{-4}	0.21	1	(0,0,**3**)
overall distribution over clusters 0, 8, **9**:					(115,118,**14**)

Table 7.2: Interesting subgroups in Cluster 9 of $\mathcal{C}^*_{10}$.

Id.	Description	p-value	$\frac{\lvert sg\cap cl\rvert}{\lvert cl\rvert}$	$\frac{\lvert sg\cap cl\rvert}{\lvert sg\rvert}$	Class distr.
A21	BNT: 14-15 & CDT: 1-2	2.10^{-4}	0.6	0.56	(4,**40**,13,1,3,6,5)
A22	age: 0-54	$< 10^{-4}$	0.39	0.79	(3,**27**,1,0,0,2,1)
A23	word list recall no: 10 & CERAD-sum: 76-100	$< 10^{-4}$	0.42	0.7	(1,**29**,5,0,0,0,2)
A24	constructional praxis: 10-11 & constructional praxis recall:10-11	$< 10^{-4}$	0.42	0.78	(1,**29**,7,0,1,0,3)
A26	BNT: 14-15 & MMSE: 26-29 & verbal fluency: $\geq$ 20	$< 10^{-4}$	0.33	0.89	(0,**23**,3,0,0,0,0)
overall distr. over clusters 1, **4**, 5, 6, 7, 11, 14:					(22,**69**,58,17,13,37,29)

Table 7.3: Interesting subgroups in Cluster 4 of $\mathcal{C}^*_{16}$.

7.3.4 Subgroup Discovery on $\mathcal{C}^*_{16}$

Analogously to the filtering procedure for $\mathcal{C}^*_{10}$, we only kept seven clusters for the further analysis of $\mathcal{C}^*_{16}$: Clusters 1, 4, 5, 6, 7, 11, and 14, containing (22, 69, 58, 17, 13, 37, 29) images (initially (42, 105, 104, 28, 40, 61, 48)). Figure 7.4 shows those clusters and those having more than five images before the filtering steps (Cluster 3 and Cluster 8).

In this clustering we find a cluster (Cluster 4) that resembles the mean image of the healthy controls. Since no hypometabolism is visible, medical experts interpreted it as representing a group of almost healthy patients. Table 7.3 shows a subset of the interesting subgroups discovered for Cluster 4. Groups of young patients (A22) and groups of patients with very good test results (A21, A23) were discovered. So, the clustering identified a relatively healthy group that was supported by both subgroup discovery and experts.

Cluster 7 describes patients that have a huge deficit in the frontotemporal metabolism. This leads to the conclusion that they are not affected by Alzheimer's disease, but by some other form of dementia. Furthermore, experts assumed no significant reduction of *CERAD constructional praxis*, while *CERAD verbal fluency* is highly impaired as well as scores in *CDR*. Significant subgroups that show the impairment were found for clustering with $k = 16$ and $k = 10$.

The mean image of Cluster 11 is the prototype of a patient with Alzheimer's disease, which is confirmed by the subgroups displayed in Table 7.4. 78.4% of the patients in this cluster have the disease, which is visible in the mean image through the reduction of metabolism in the temporoparietal cortex. In fact, this is not the only cluster with a

Id.	**Description**	p**-value**	$\frac{\lvert sg \cap cl \rvert}{\lvert cl \rvert}$	$\frac{\lvert sg \cap cl \rvert}{\lvert sg \rvert}$	**Class distr.**
A101	ICD: Alzheimer's	$< 10^{-4}$	0.78	0.33	(11,5,16,12,3,**29**,13)
A102	BNT: 14-15 & gender: m & ICD: Alzheimer's	$< 10^{-4}$	0.32	0.71	(0,0,3,1,1,**12**,0)
A106	age: 55-59 & ICD: Alzheimer's	5.10^{-4}	0.19	0.58	(1,1,0,3,0,**7**,0)
A107	age: 65-69 & ICD: Alzheimer's	10^{-4}	0.3	0.58	(3,1,3,1,0,**11**,0)
overall distr. over clusters 1, 4, 5, 6, 7, **11**, 14:					(22,69,58,17,13,**37**,29)

Table 7.4: Interesting subgroups in Cluster 11 of $\mathcal{C}^*_{16}$.

high ratio of patients with Alzheimer's disease. In Cluster 6, 70% of the patients suffer from the disease. This was confirmed by subgroups containing an Alzheimer's diagnosis. Contrary to Cluster 11, in Cluster 6 also the frontal metabolism is reduced. This is reflected by worse (higher) results of CDR.

Regarding the subgroups found for Cluster 14 (Table 7.5), it seems that this cluster describes elderly women (A124, A129, A133) with low test results and therefore a similar state of dementia. Surprisingly, there is a group of men with high MMSE scores (A126). Although this subgroup is not significant, it is highly interesting. It indicates that men with the same metabolic patterns as women have a less impaired cognitive ability. Further investigations (Table 7.6) showed that men and women are in the same age group, but men do have slightly better overall results in the most important psychological tests. Although female and male patients fall into the same cluster (based on brain metabolism), they apparently differ in their cognitive abilities. This finding may possibly be explained by the hypothesis of cognitive reserve, which postulates that some individuals can somehow offset the symptoms of neurodegeneration. Although the neurobiological substrate is still unknown, the higher neuron count in men might be associated with higher reserve. To show the same symptoms of dementia as women, men have to suffer from a larger loss of cells. Another factor discussed in dementia research [75] is the education level, which is also higher among the men in this cluster, compared to the women. Here we now see an example of the power of subgroups. The method allows not only to describe sets of instances, but also to bring up groups with unusual and therefore interesting features, which cannot be found by simple correlation studies. Altogether, we can conclude that the clustering with $k = 16$ produces more meaningful clusters than the clustering with $k = 10$. Even though all subgroups displayed are statistically significant (see the p-values), $\mathcal{C}^*_{16}$ differentiates more accurately. For example, it detects an "Alzheimer cluster" (Cluster 11), whereas the distribution of patients with Alzheimer's disease in $\mathcal{C}^*_{10}$ is (39, 48, 5). Therefore, none of the clusters in $\mathcal{C}^*_{10}$ shows a preference of being a definite "Alzheimer cluster".

Id.	**Description**	p**-value**	$\frac{\lvert sg \cap cl \rvert}{\lvert cl \rvert}$	$\frac{\lvert sg \cap cl \rvert}{\lvert sg \rvert}$	**Class distr.**
A121	graduation < high school & gender: f	2.10^{-1}	0.52	0.24	(4,17,15,3,1,8,**15**)
A124	ICD: Alzheimer's & personal: 0-0.5 & gender: f & age: 70-73	2.10^{-3}	0.17	0.83	(0,0,0,1,0,0,**5**)
A126	wordlist recall: 7-10 & gender: m & age: 65-69 & MMSE: 30	2.10^{-1}	0.07	1	(0,0,0,0,0,0,**2**)
A127	other tests: MR & graduation < high school & age: 74-77	7.10^{-3}	0.24	0.54	(1,1,3,1,0,0,**7**)
A129	gender: f & MMSE: 11-20 & age: 70-73	10^{-2}	0.17	0.71	(0,0,1,0,0,1,**5**)
A133	gender: f & CERAD-sum: 47-57 & age: $\geq$ 78	10^{-2}	0.14	1	(0,0,0,0,0,0,**4**)
overall distr. over clusters 1, 4, 5, 6, 7, 11, **14**:					(22,69,58,17,13,37,**29**)

Table 7.5: Interesting subgroups in Cluster 14 of $\mathcal{C}^*_{16}$.

	Women		**Men**	
		n		n
Age	75	19	70	10
MMSE	23.26	19	25.22	9
global	0.97	15	0.8	5
CERAD-Sum	53.64	14	62.5	8
CDT	3.93	15	3.66	9

Table 7.6: Comparison of the gender distribution in Cluster 14 of $\mathcal{C}^*_{16}$.

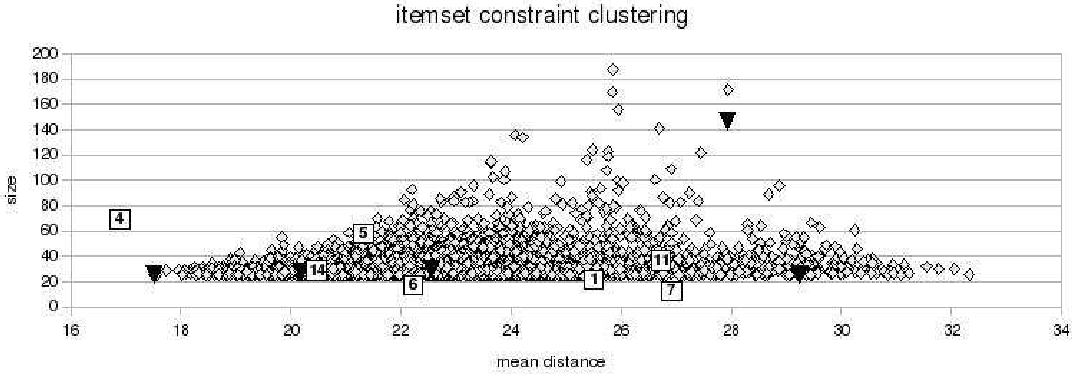


Figure 7.9: Mean within-cluster distance (x-axis) and support (y-axis) of resulting itemsets (clusters) with a minimum relative support of 0.1.

7.4 Comparison with Constrained Clustering

Another possible approach is to apply methods from constraint-based clustering (e.g., by Sese *et al.* [81]), such that only clusters constrained by descriptions of non-image variables are considered. This is similar to the usual approach in medicine: select a subset of patients fulfilling specific predefined criteria and compare the images associated with those patients. Automating this process results in determining frequent itemsets based on the structured non-image data. This can easily be achieved with the APRIORI algorithm [92]. First we select a subset of patients covered by one frequent itemset. Then we evaluate the similarity of their PETscans by the mean of the pairwise weighted Euclidean distance. In this way, we obtain a mean distance for each itemset.

The grey diamonds in Figure 7.9 represent the discovered itemsets, where the x-axis corresponds to their mean distance and the y-axis to the size of the subset covered by the itemset. We found 5,858 frequent itemsets for a *minsupport* of 0.1. For comparison, the white squares represent the large clusters from clustering $\mathcal{C}^*_{16}$. We can see that Cluster 4 (similar to healthy controls) has a lower mean distance (higher similarity of brain activity) than any of the itemset constraint groups. However, there is a large number of itemsets with more homogeneous PET scans than Cluster 11 and Cluster 7. This is a result of the clustering approach, where the task is to find a set of disjoint clusters whose union covers the entire example set. Therefore, it is necessary to also put patients with dissimilar PET scans into a cluster.

The clustering task constrained to itemsets means finding a set of disjoint frequent itemsets that together cover all examples. The straight-forward approach is to do that in a greedy way: we determine the itemset with the most homogeneous PET scans. Then we determine the next most homogeneous itemset which does not overlap with the previously determined itemsets and so on. When no more disjoint itemset can be found, we combine the yet uncovered, remaining examples into a final set. The black triangles in Figure 7.9 represent the set of disjoint itemsets. It shows that two of the disjoint itemsets have higher mean distance than the clusters in clustering $\mathcal{C}^*_{16}$. These are the sets that were selected last in the greedy algorithm. We therefore want to consider different algorithm of selecting a disjoint set of frequent itemsets. We do that by modeling this task as an integer-linear program in the next chapter.

In summary, we can say that itemset-constraint-clustering provides physicians with groups of patients that share similar psychological and similar metabolic patterns (itemsets with low mean distance). Additionally, it finds groups with similar psychological patterns and higher variation in brain activity (itemsets with high mean distance). However, this approach is not able to detect a complete set of patients with similar brain activity. On the other hand, clustering of PET scans finds groups of patients with similar brain activity. That means our approach can combine all patients with a particular metabolic pattern, even patients that differ in their psychological features. Therefore, itemset-constraint clustering seems to be adapted for specific queries, whereas our approach provides a general correlation of image and non-image data. Moreover, we automatically create an overview of the status of disease and provide hypotheses to be further validated by medical experts.

7.5 Discussion

As they reflect, in some sense, the “ground truth” of the state of a brain, we chose PET images as the starting point of our analysis. The differences between different states are then explained by non-image variables. The presented approach uses k-Medoids and RSD to achieve those tasks, but we also considered and tested other methods for clustering and correlation analysis. Due to the high dimensionality of the data, we also tested the subspace clustering method PreDeCon [16], but could not identify appropriate parameter settings to obtain significant results. As a further test, we clustered the images with hierarchical agglomerative clustering (complete linkage). This method also found the outliers mentioned above and sorted the relatively healthy patients into one cluster. However, although the two clustering methods produced largely similar results, they differed in the partition of unspecific clusters, which may be explained by the “myopia” of the hierarchical clustering scheme. To establish a baseline for the subgroup discovery, we performed a simple correlation analysis based on individual variables. Our experiments confirmed that this is clearly not sufficient, as many of the more complex subgroups cannot be detected in this way.

From a medical point of view, we showed that it is possible to obtain meaningful clusters of PET scan images. Medical experts could identify different patterns of disease, and confirmed that the resulting findings (clusters, outliers and subgroups) have medical novelty and significance. As mentioned above, the new approach of reversing the direction of the analysis, i.e., starting with the PET scan images, is considered new in this area.

7.6 Related Work

Data mining in the context of dementia and Alzheimer's disease can be roughly categorized into three categories: First, machine learning approaches for the improvement of differential diagnosis. Second, data mining techniques applied to brain imaging data. Third, transcriptomics and proteomics analysis of dementia and Alzheimer's disease.

One of the earliest work from the first category deals with the prediction of the type of dementia (Alzheimer's disease, vascular dementia, or other) from a small set of demographic variables and the total scores from clinical tests [62]. The latter variables included measures of category fluency, letter fluency, delayed free recall and recognition, simple and complex attention span, visual-constructional abilities, and object naming. Similarly, Corani *et al.* [27] predicted different types of dementia (Alzheimer's disease, dementia with Lewy bodies, Parkinson's disease with dementia and vascular dementia) from cognitive profiles based on the Cognitive Drug Research (CoDR) system. CoDR consists of a series of computerized tests (tasks), which assess some cognitive faculties of the patient, such as memory, attention, and reaction times. The results from those tests together constitute the cognitive profile of a patient.

Work from the second category includes the one by Fung *et al.* [37] and Megalooikonomou *et al.* [63]. Fung *et al.* classify Alzheimer's patients based on their SPECT images. Megalooikonomou *et al.* give a survey of data mining techniques applied to brain imaging data. Clustering is applied to find groups of inter-related voxels, not to find groups of images. To the best of our knowledge, clustering methods have not been applied to whole PET scans before.

The third category includes a study of gene expression changes in patients with Alzheimer's disease [88] and a study aiming for the discovery of Alzheimer-relevant proteins [24].

The work presented in this chapter differs from previous work in its combined analysis of PET images and clinical variables. Similar to work in the first category, the results are mainly useful for diagnostic purposes. Although biological information in the form of transcriptomics or proteomics data is not yet used in our approach, it is easy to incorporate it in the subgroup descriptions or even as the target for subgroup discovery.

7.7 Conclusion

In the chapter, we introduced a new and challenging problem for data mining research: correlating large databases of PET scans with structured patient data. The goal of the work was not to develop completely new methods, but to show that current data mining methods like RSD are able to solve this large (initially 200 GB) and complex (image data and 11 relations) problem and produce valid and relevant results. The task itself is critical to gain a better understanding of various forms of dementia. The presented approach aims for more completeness than previous methods. To do so, we first identified clusters of PET scans sharing similar features in brain metabolism. In the second step, we explained the differences and commonalities among those clusters in terms of clinical and demographic variables. To validate the results, we computed p-values of the clusterings and interpreted the clusters and subgroup descriptions in the light of domain knowledge. To the best of knowledge, this type of analysis has not been done before. One of the subproblems, the clustering of whole PET scans (not voxels), also has not been addressed in the literature before.

In future work, we are planning to further improve the quality of the clusterings by taking into account more advanced features (e.g., brain regions) and by developing more advanced methods specifically for high-dimensional PET images. Moreover, standard algorithms for subgroup discovery suffer from similar problems as algorithms for pattern mining and association rule mining [74]. For instance, it is necessary to filter interdependent results, as the refinement (specialization) of an "interesting" subgroup is likely to produce another "interesting" subgroup. Another limiting factor of the subgroup discovery approach is the incompleteness of the psychological data. Finally, the integration of gene expression and proteomic data could aid in the formation of mechanistic hypotheses.

Another possible approach introduced in Section 7.4 is to apply methods from constraint-based clustering, such that only clusters constrained by descriptions of non-image variables are considered. We will further investigate this task in the next chapter.

In summary, we believe that explaining medical images in terms of other variables (patient records, demographic information, etc.) is a challenging new and rewarding task for data mining research.

8 Constrained Clustering

In this chapter, we address the problem of building a clustering as a subset of a (possibly large) set of candidate clusters under user-defined constraints. In contrast to most approaches to constrained clustering, we do not constrain the way observations can be grouped into clusters, but the way candidate clusters can be combined into suitable clusterings. The constraints may concern the type of clustering (e.g., complete clusterings, overlapping or encompassing clusters) and the composition of clusterings (e.g., certain clusters excluding others). We show that these constraints can be translated into integer linear programs, which can be solved by standard optimization packages. Our experiments with benchmark and real-world data investigates the quality of the clusterings and the running times depending on a variety of parameters.

8.1 Motivation

Constraint-based mining approaches aim for the incorporation of domain knowledge and user preferences into the process of knowledge discovery [19]. This is mostly supported by inductive query languages [20, 76, 72, 34, 18]. One of the most prominent and important instances of constraint-based mining is constrained clustering. Since its introduction (incorporating pairwise constraints into k-Means [87]), constrained clustering has been extended to various types of constraints and clustering methods (compare Chapter 3.3). Most of the approaches constrain the way observations can be grouped into clusters, i.e., they focus on *building clusters* under constraints. In this chapter, we consider a different problem, namely that of *building clusterings* from a (possibly large) set of candidate clusters under constraints. In other words, we address the following problem: Given a set of candidate clusters, find a subset of clusters that satisfies user-defined constraints and optimizes a score function reflecting the quality of a clustering. Clearly, both approaches are not mutually exclusive, but just represent different aspects of finding a good clustering under constraints. In fact, the problem of constructing suitable clusterings requires a suitable set of cluster candidates, which can be the result of, e.g., constrained clustering under pairwise constraints [87] or itemset classified clustering [82].

The process of building suitable clusterings can be constrained by the user in various ways: The constraints may concern the completeness of a clustering, the disjointness of clusters, or they may concern the number of times examples are covered by clusters. Moreover, some clusters may preclude others in a clustering, or clusters may require others. Constraints of the latter type can be formulated as logical formulae. The quality of a clustering to be optimized can then be defined as the mean quality of the clusters, their median quality, or their minimum quality. We present a set of possible constraints along those lines and shows how they can be mapped onto integer linear program models.

In this way, users can obtain tailor-made clusterings without being concerned with the technical details, much in the spirit of constraint-based mining and inductive query languages in general [19, 20]. Doing so, it is also possible to take advantage of a huge body of literature on the subject and advanced optimization tools.

This chapter is organized as follows: In Section 8.2, we present the different types of constraints that can be considered. Section 8.3 explains how those constraints can be translated into an integer linear model. In Section 8.4, we present experimental results with the approach. Then, we discuss related work (Section 8.5) and come to our conclusions (Section 8.6).

8.2 Constrained Clustering

In this section, we present the possible constraints that can be applied to a clustering. First, we have to introduce some notation: Let $X = \{e_1, ..., e_m\}$ be a set of examples, and B a set of base clusters (i.e., cluster candidates potentially to be included in a clustering). Furthermore, the number of examples is denoted by $m = |X|$, and the number of base clusters by $n = |B|$. Then a set of clusters $C = \{C_1, ..., C_k\} \subseteq B$ denotes a clustering. Moreover, we are given an objective function $f : 2^B \to \mathbb{R}$ which scores a given clustering according to its quality. Finally, we are given constraints $\Phi(C)$ that restrict the admissible subsets of B, either with respect to the sets of instances (e.g., whether they are overlapping or encompassing one another) or with respect to known interrelationships between the clusters (e.g., cluster C_1 precludes cluster C_2 in a clustering).

The overall goal is then to find a clustering $C \subseteq B$ satisfying the constraints $\Phi(C)$ and optimizing the objective function f.

Next, we discuss the different types of constraints on clusterings. First, we present *set-level constraints*, i.e., constraints in the form of logical formulae that control the clusters that can go into a clustering depending on other clusters. Second, we present *clustering constraints*, that is, constraints that determine the form of a clustering, for instance, whether the clusters are allowed to overlap. Third, three different *optimization constraints* [34] will be introduced, which determine the objective function to be optimized.

8.2.1 Set-Level Constraints

In the following, let a *literal* Lit_i be either a constraint $C_j \in C$ or $C_j \notin C$. For convenience, we will call $C_j \in C$ an *unnegated literal* and $C_j \notin C$ a *negated literal.* Then we can have three types of constraints:

- Conjunctive constraints $setConstraint(C, Lit_1 \wedge Lit_2 \wedge ... \wedge Lit_l)$: This type of constraint ensures that certain clusters are included or excluded from a clustering. Clusters referred to by unnegated literals have to be included, clusters referred to by negated literals have to be excluded.

- Disjunctive constraints $setConstraint(C, Lit_1 \vee Lit_2 \vee ... \vee Lit_l)$: This type of constraint ensures that at least one of the conditions holds for a clustering. If all literals are unnegated, for instance, it is possible to state that at least one of the listed clusters has to participate in a clustering.
- Clausal constraints $setConstraint(C, Lit_1 \wedge Lit_2 \wedge ... \wedge Lit_{l-1} \rightarrow Lit_l)$: This type of constraint states that if all conditions from the left-hand side are satisfied, then also the condition on the right-hand side has to be satisfied. For instance, it is possible to say that the inclusion of one cluster has to imply the inclusion of another (e.g., $C_i \in C \rightarrow C_j \in C$) or that one clusters makes the inclusion of another impossible (e.g., $C_i \in C \rightarrow C_j \notin C$).

These constraints can be translated easily into linear constraints (see Section 8.3). Arbitrary Boolean formulae could, in principle, be supported as well. However, they would require the definition of new variables, thus complicating the definition of the optimization problem. Note that *instance-level constraints* in the style of Wagstaff *et al.* [87], *must-link* and *cannot-link*, can easily be taken into account as well: However, here the constraints would effectively reduce the set of clusters that are put into the set of candidate clusters B. If we enforced those constraints on some set of candidate clusters B, then we would discard those individual clusters not satisfying the constraints, and restrict B to some subset $B' \subseteq B$ in the process.

Also note that the set of base clusters B can be the result of other data mining operations. For instance, consider the case of *itemset classified clustering* [82], where the potential clusters in one feature space (view) are restricted to those that can be described by frequent itemsets in another feature space (view). This is the setting that will be explored in Section 8.4 on experimental results.

8.2.2 Clustering Constraints

This section describes constraints that determine the basic characteristics of clusterings:

- $completeness(C, minCompl, maxCompl)$: This constraint determines the degree of completeness of a clustering. More formally, it ensures that for clustering C, it holds that $minCompl \leq \frac{|\cup_{C_j \in C} C_j|}{|X|} \leq maxCompl$.
- $overlap(C, minOverlap, maxOverlap)$: This constraint determines the allowed degree of overlap between clusters of a clustering. Let $coverage(e_i, C)$ be a function determining the number of clusters in a clustering C containing example e_i: $coverage(e_i, C) = |\{C_j | C_j \in C \wedge e_i \in C_j\}|$. Furthermore, let $numberOverlaps(C) = \sum_{e_i \in X,\ coverage(e_i, C) > 1} (coverage(e_i, C) - 1)$ be a function counting the number of times instances are covered more than once (with multiple overlaps of one instance counted multiple times). Then the constraint $overlap(C, minOverlap, maxOverlap)$ is satisfied if $minOverlap \leq \frac{numberOverlaps(C)}{|X|} \leq maxOverlap$.[1]

[1] This is related to the notion of disjointness of clusters, but minimum overlap and maximum overlap seem more intuitive than minimum disjointness and maximum disjointness.

- $encompassing(C, Flag)$: This constraint determines whether the clusters are allowed to encompass each other, i.e., whether the clusters are allowed to form a hierarchy. If $Flag = no$, then it holds that there is no pair $C_i, C_j \in C$ such that $C_i \subset C_j$.
- $numberClusters(C, minK, maxK)$: This constraint restricts the number of clusters that can be part of a solution: $minK \leq |C| \leq maxK$.
- $exampleCoverage(C, minCoverage, maxCoverage)$: This constraint limits the number of times an example can be covered by clusters. Formally, it holds that for each $e_i \in X : minCoverage \leq coverage(e_i, C) \leq maxCoverage$.

8.2.3 Optimization Constraints

This section introduces the optimization constraints [34] used in our approach. The quality of a clustering is defined as an aggregate over the qualities of its clusters.

- $maxMeanQuality(C)$: This constraint implies that the mean quality of the clusters contained in a clustering is optimized.
- $maxMinQuality(C)$: This constraint implies that the minimum quality of the clusters contained in a clustering is optimized.
- $maxMedianQuality(C)$: This constraint implies that the median quality of the clusters contained in a clustering is optimized.

Those optimization constraints are just the most basic ones that are conceivable. In fact, it is easy to combine the quality of a clustering with any of the above clustering constraints, for instance, not excluding overlapping clusters, but penalizing too much overlap. The same is possible for completeness (penalizing lack of completeness). Note that if cluster quality is defined as within-cluster distance, it is necessary to invert this quantity (e.g., by changing the sign) for our purposes. When translated into an integer linear model, the optimization constraints determine the objective function used.

8.2.4 Combining Constraints

Given the constraints introduced above, it is now possible to combine them in queries for clusterings. More precisely, a query can now be formed by a pair $(\Phi, f())$, where Φ is a logical conjunction of set-level and clustering constraints, and $f()$ is one of the three optimization constraints from the previous section.

As an example, consider the following query:

$$\begin{aligned} q = (&numberClusters(C, 2, 5) \wedge overlap(C, 0.1, 0.2) \wedge completeness(C, 0.9, 1.0) \wedge \\ &setConstraint(C, C_1 \in C \rightarrow C_2 \in C) \wedge setConstraint(C, C_1 \in C \rightarrow C_3 \notin C), \\ &maxMeanQuality(C)) \end{aligned}$$

It aims to find a clustering containing between 2 and 5 clusters, with an allowed overlap between 0.1 and 0.2, a desired completeness between 0.9 and 1.0, and such that the inclusion of cluster C_1 implies the inclusion of C_2 and the exclusion of C_3. The possible clusterings are optimized with respect to their mean cluster quality.

8.3 Method

In this section, we describe how to translate the introduced constraints into linear constraints. In Section 8.3.1 and 8.3.2 we focus on the clustering constraints restricted to the optimization constraint $maxMeanQuality(C)$, in Section 8.3.3 we handle alternative optimization constraints and in Section 8.3.4 we present the translation of set-level constraints. The constraints are used to form an integer linear program, which can then be solved by any package for ILP optimization.

First, we define an $m \times n$-matrix A (m being the number of examples and n being the number of base clusters as introduced above) with:

$$a_{ij} = \begin{cases} 1 & \text{if cluster } C_j \text{ contains example } e_i, \\ 0 & \text{otherwise,} \end{cases} \tag{8.1}$$

and $w \in \mathbb{R}^n$, where w_j is the *within-cluster distance*, which is defined as the mean of the pairwise distances between the examples covered by the cluster C_j. Note that minimizing the within-cluster distance is equivalent to maximizing the between-cluster distance for a fixed k. In our setting, k varies between $minK$ and $maxK$. As our clustering approach is strongly constrained by the given set of base clusters, there is no or only little bias towards a k near $maxK$. For instance, smaller base clusters may not provide the required completeness, thus, a solution with larger base clusters and a smaller k may be preferred.

Figure 8.1 shows an example for $m = 7$ examples and $n = 5$ candidate clusters. The matrix A and the vector w are displayed in Table 8.1. Here, w_j is the mean of the pairwise Euclidian distances between the examples covered by cluster C_j.

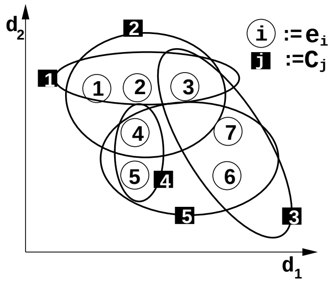


Figure 8.1: Problem with 7 examples and 5 base clusters

A:	C_1	C_2	C_3	C_4	C_5
e_1	1	1	0	0	0
e_2	1	1	0	0	0
e_3	1	1	1	0	0
e_4	0	1	0	1	1
e_5	0	0	0	1	1
e_6	0	0	1	0	1
e_7	0	0	1	0	1

w:	1.33	1.3	1.55	1	1.75

Table 8.1: Matrix A and vector w for the example introduced in Fig. 8.1

minCompl	maxOverlap	selected sets	mean(w)
1	0	$\{C_1, C_5\}$	1.54
6/7	0	$\{C_1, C_5\}$	1.54
5/7	0	$\{C_1, C_4\}$	1.17
1	1/7	$\{C_1, C_3, C_4\}$	1.29
1	2/7	$\{C_2, C_3, C_4\}$	1.28
5/7	2/7	$\{C_2, C_4\}$	1.15

Table 8.2: Optimal clusterings for the example introduced in Fig. 8.1

maximize		$\frac{1}{k}(w_{max} - w)^T x$
subject to	(i)	$Ax \leq \mathbf{1}$
	(ii)	$Ax \geq y$
	(iii)	$\mathbf{1}^T x = k$
	(iv)	$\mathbf{1}^T y \geq m \cdot minCompl$
	(v)	$x \in \{0,1\}^n$
	(vi)	$y \in \{0,1\}^m$

Table 8.3: Optimization task for determining the optimal disjoint clustering

8.3.1 Modeling Clustering Constraints: Disjoint Clustering

Let our objective be to determine the disjoint clustering C with the minimal mean within-cluster distance, i.e., C has to satisfy

$$(completeness(C, minCompl, 1) \wedge overlap(C, 0, 0), maxMeanQuality(C)).$$

This is also known as the Weighted Set Packing Problem [38], which is NP-complete. This task can be defined as the optimization problem shown in Table 8.3.

The goal is to minimize the mean of the within-cluster distance (w) over all k selected clusters, which is equivalent to maximizing the mean of the inner cluster similarities ($w_{max} - w$) of all k selected clusters, where $w_{max} := \max_j w_j$.

Since we would like to formulate a linear program, it is not possible to optimize over the variable k that appears in the denominator of the objective function. However, we can keep the problem linear by treating k as a constant and resolve the optimization problem with varying values for k.

We introduce a vector x expressing which clusters are selected (i),(v):

$$x_j = \begin{cases} 1 & \text{if cluster } C_j \text{ is selected,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.2}$$

The vector y contains information about which examples are covered by the selected clusters: If $y_i = 1$ then e_i is covered by a selected cluster (ii).[2]

The clustering is further subject to the constraints that each example must not be covered by more than one clustering (i) and that at least $m \cdot minCompl$ examples have to be covered (iv).

For the example in Figure 8.1, we obtain the solutions presented in the first three rows of Table 8.2.

[2] Note: if we set $minCompl < 1$, it is possible that $y_i = 0$ for some example e_i, even though e_i is covered by the selected clusters. However, this does not affect the solution.

$$
\begin{array}{lrl}
\text{maximize} & & \frac{1}{k}(w_{max} - w)^T x \\
\hline
\text{subject to} & \text{(i)} & Ax = y \\
 & \text{(ii)} & \mathbf{1}^T x = k \\
 & \text{(iii)} & z \geq \mathbf{1} - y \\
 & \text{(iv)} & \mathbf{1}^T z \leq (m - m \cdot minCompl) \\
 & \text{(v)} & v \geq y - \mathbf{1} \\
 & \text{(vi)} & \mathbf{1}^T v \leq n \cdot maxOverlap \\
 & \text{(vii)} & x \in \{0,1\}^n \\
 & \text{(viii)} & y \in \mathbb{N}_0^m \\
 & \text{(ix)} & v \in \mathbb{N}_0^m \\
 & \text{(x)} & z \in \{0,1\}^m
\end{array}
$$

Table 8.4: Optimization task for determining the optimal clustering with up to *maxOverlap* multiply covered examples

8.3.2 Modeling Clustering Constraints: Clustering with Overlaps

We can relax the constraint of disjointness by allowing some of the selected clusters to overlap. This means that some examples can be covered by more than one cluster: C has to satisfy

$$(completeness(C, minCompl, 1) \wedge overlap(C, 0, maxOverlap), maxMeanQuality(C)).$$

This can be realized by the optimization task in Table 8.4. The goal is still to maximize the mean of the inner cluster similarity of the k selected sets. Again, we have the constraint that at least $m \cdot minCompl$ examples have to be covered (iv). In this setting we allow that some examples can be covered by more than one set. We restrict the number of allowed overlaps to *maxOverlap* (vi). This yields an integer-valued vector y, where y_i = number of sets that cover example e_i ($y(i) = coverage(e_i, C)$). Furthermore, we need the vector v, where v_i = number of overlaps of the example e_i.[3] To model the constraint that demands at least *minCompl* examples to be covered, we introduce a vector z such that if $z_i = 0$ then example e_i is covered.

For the example in Figure 8.1, we obtain the solutions presented in the lower part of Table 8.2 where $maxOverlap > 0$.

8.3.3 Modeling Optimization Constraints

So far we have shown integer linear models that determine the optimal clustering C with respect to $maxMeanQuality(C)$. In this section we will show how to model the

[3] If the optimal C contains less overlaps than $n \cdot maxOverlap$, v_i may take higher values than the actual number of overlaps of e_i. (v) However, this does not affect the solution.

optimization constraints $maxMinQuality(C)$ and $maxMedianQuality(C)$.

The optimization task in Table 8.4 has to be modified in the following way:

maxMinQuality(C)

Instead of maximizing the mean $\frac{1}{k}(w_{max}-w)^T x$, the aim is now to maximize the objective function d_{min}, that is the lowest inner cluster similarity. For this purpose, we introduce the additional constraint $d_{min} \leq w_{max} - w_j x_j$ for each j, making sure that d_{min} takes the intended value. With this objective, we can remove the constraint $\mathbf{1}^T x = k$.

maxMedianQuality(C)

Here, the objective is to maximize d_{med}, which means to maximize the median quality. Again, we can remove the constraint $\mathbf{1}^T x = k$. We need to introduce the following additional constraints:

(xi) $x_l \in \{0,1\}^n$
(xii) $x_r \in \{0,1\}^n$
(xiii) $x = x_l + x_r$
(xiv) $x_l + x_r \leq \mathbf{1}$
(xv) $\mathbf{1}^T x_r - \mathbf{1}^T x_l \leq 1$
(xvi) $\mathbf{1}^T x_l - \mathbf{1}^T x_r \leq 0$
(xvii) $(w_{max} - w_j)x_{l_j} \leq d_{med},\ \forall j$
(xviii) $d_{med} \leq w_{max} - w_j x_{r_j},\ \forall j$

The intuitive explanation for the introduced vectors x_l and x_r is as follows. To determine the median inner cluster similarity we partition the selected clusters into two sets whose cardinalities differ by at most 1 (constraints (xv) and (xvi)). The clusters that have a lower quality than the median-cluster quality are those having an x_l value of 1 (constraint (xvii)), and the clusters that have a higher quality than the median-cluster quality are those having an x_r value of 1 (constraint (xviii)).

Note that for the case of an even number of sets, this model maximizes the upper median, i.e., the set at the $(\frac{k}{2}+1)$th position. If the goal is to maximize the lower median, i.e., the set at the $\frac{k}{2}$th position, we need to modify the constraints (xv) and (xvi) to:

(xv) $\mathbf{1}^T x_r - \mathbf{1}^T x_l \geq 1$
(xvi) $\mathbf{1}^T x_r - \mathbf{1}^T x_l \leq 2$

Number of constraints

The introduced models (including the overlap and the completeness constraints) consist of $7m+2n+3$ constraints for $maxMeanQuality(C)$, $7m+3n+2$ constraints for *maxMinQuality(C)*, and $7m + 10n + 4$ constraints for *maxMedianQuality(C)*. Thus the number of constraints directly depends on the dimensions of the input matrix A. However, the number of constraints is not the only factor that impacts the computational time. Moreover, the time to solve the problem depends on the number k of desired clusters and the parameter values $minCompl$ and $maxOverlap$. In which way these parameters influence the runtime mainly depends on the structure of the underlying data set and the selection of base clusters. Compare our experiments on the scalability of the approch in Section 8.4.1.

8.3.4 Modeling Set-Level Constraints

Finally, we explain how set-level constraints can be dealt with, and based on this, how the $encompassing(C, Flag)$ constraint can be solved. For convenience, we define a transformation operator τ on literals, which gives $\tau(C_j \in C) = x_j$ in case of unnegated and $\tau(C_j \notin C) = (1 - x_j)$ in case of negated literals.

- Conjunctive constraints $setConstraint(C, Lit_1 \wedge Lit_2 \wedge ... \wedge Lit_l)$. This states that certain clusters have to be included or cannot be included in a clustering. These constraints can directly be transformed into equality constraints of the form $x_j = 1$ for $C_j \in C$ and $x_j = 0$ for $C_j \notin C$.
- Disjunctive constraints $setConstraint(C, Lit_1 \vee Lit_2 \vee ... \vee Lit_l)$. This gives rise to an additional linear constraint of the following form: $\sum_{i=1}^{l} \tau(Lit_i) \geq 1$.
- Clausal constraints $setConstraint(C, Lit_1 \wedge Lit_2 \wedge ... \wedge Lit_{l-1} \rightarrow Lit_l)$. This gives rise to an additional linear constraint $\tau(Lit_l) - (\sum_{i=1}^{l-1} \tau(Lit_i)) \geq 2 - l$. For instance, $setConstraint(C, C_1 \in C \wedge C_2 \notin C \rightarrow C_3 \in C)$ gives rise to the constraint $x_3 - x_1 - (1 - x_2) \geq -1$, i.e., $x_3 - x_1 + x_2 \geq 0$, which is only violated for $x_1 = 1, x_2 = 0, x_3 = 0$.

Using these set-level constraints, it is now possible to solve the $encompassing(C, Flag)$ constraint: If $Flag = no$, then for each $C_i \in B$ and $C_j \in B$ with $C_i \subset C_j$, the following set constraint has to be set: $setConstraint(C, C_i \notin C \vee C_j \notin C)$. In other words, it will be translated into a linear constraint $(1 - x_i) + (1 - x_j) \geq 1$, i.e., $x_i + x_j \leq 1$.

8.4 Experiments and Results

We implemented the two linear models of Table 8.5 and Table 8.4 and tested their performance on two datasets. For optimization, we use the Xpress-Optimizer [29], which combines common methods, such as the simplex method, cutting plane methods, and branch and bound algorithms [79].

For the first batch of experiments we use the dataset on dementia patients provided by the psychiatry and nuclear medicine departments of Klinikum rechts der Isar of Technische Universität München (see Section 2.2.2). Recalling from Section 2.2.2, the dataset consists of two types of data: structured data (demographic information, clinical data, including neuropsychological test results) and image data (PET scans showing the patient's cerebral metabolism). We include 257 data records belonging to patients with PET scans and psychological and demographic values.

minSupport	0.05	0.1	0.2	0.3
$n =$ #itemsets	68,460	5,447	326	62
#possible pairs	$2.3 \cdot 10^9$	$1.5 \cdot 10^7$	$5.3 \cdot 10^4$	180
# disjoint pairs		$1.2 \cdot 10^6$	$1.8 \cdot 10^3$	40
# clusters that a cluster overlaps				
min	4,079	1,451	234	47
max	68,477	5,447	326	62
mean	56,435	5,014	314	59
median	60,921	5,191	319	59
# clusters that cover an example				
min	27	21	13	7
max	49,781	4,221	253	50
mean	46,933	728	85	24
median	16,285	372	73	23

Table 8.5: Statistics of frequent itemsets on the dementia data with $m = 254$ examples.

Our experimental setting is the one introduced in Section 7.4 and similar to the usual approach in medicine: select a subset of patients fulfilling specific predefined criteria and compare the images associated with those patients. Automating this process results in determining frequent itemsets based on the structured non-image data. This can easily be achieved with the APRIORI algorithm [2]. First we select a subset of patients covered by one frequent itemset. Then we evaluate the similarity of their PET scans by the mean of the pairwise weighted Euclidean distance (see equation (7.1)). In this way, we obtain a mean distance w_j for each itemset C_j. To tackle outliers, i.e., PET scans that are very distant from all other PET scans, we remove those data records, before generating itemsets. Otherwise each outlier affects the mean distance of each itemset it is covered by. For the given dataset, three outliers are removed.

Since we constrain our clustering to frequent itemsets, there is a huge number of clusters that overlap or even completely encompass other clusters, e.g., if $C_1 =$ set of examples that support item i_1 and $C_2 =$ set of examples that support items $i_1 \wedge i_2$, then $C_2 \subseteq C_1$. For a relative *minSupport* of 0.1, we obtain $5,447$ itemsets (see Table 8.5 for all related statistics). Each of them overlaps other itemsets. On average, each itemset overlaps $5,014$ other itemsets, at least $1,451$ and at most all $5,447$ sets. Each example is covered by 727.5 itemsets on average, at least by 21 and at most by $4,221$. For the distribution of other *minSupport* values see again Table 8.5.

We measure how the mean within-cluster distance[4] performs for different parameter settings. For each setting, we run the optimization task for all $k \in \{2, \ldots, \lfloor (maxOverlap + 1) \cdot n \cdot minSupport \rfloor\}$[5] and decide for the best solution of these runs. For example, for

[4] Note that our base clusters have a minimal size of $n \cdot minSupport$. Therefore, we avoid that the clustering can consist of singleton clusters with a within-cluster distance of 0.

[5] Due to the minimal cluster size, there is no solution possible for a larger k.

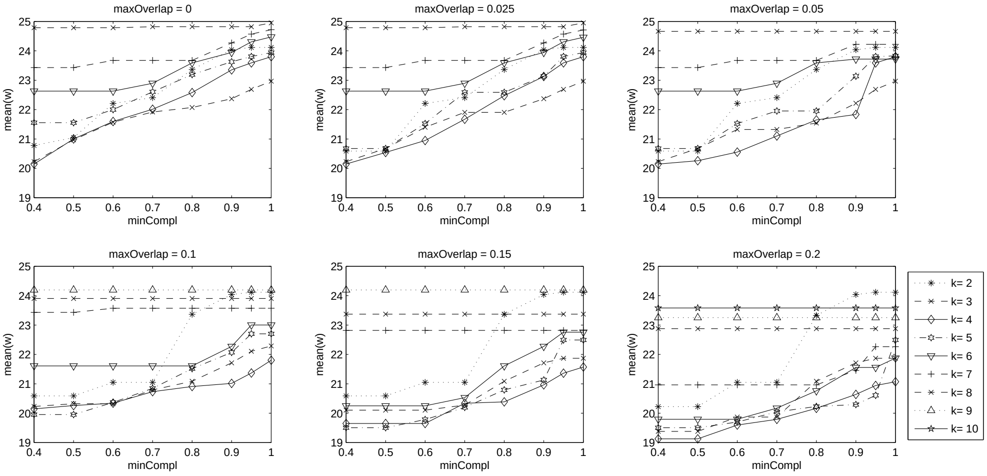


Figure 8.2: Mean within-cluster distance (y-axis) for varying *maxOverlap*, *minCompl* (x-axis), and k parameters and *minSupport* $= 0.1$.

our data with $m = 254$, a *minSupport* of 0.1 and *maxOverlap* $= 0.15$, we run it for all $k \in \{2, \ldots, 10\}$.

The results of the single runs with different ks for various parameter settings are shown in Figure 8.2. We test on six different values for *maxOverlap* $(0, 0.025, 0.05, 0.1, 0.15, 0.2)$ with *minCompl* varying from 1.0 to 0.4. For each parameter setting, we are interested in the clustering with the lowest *mean*(w), comparing the solutions of different k values. For the most parameter settings, the best clustering consists of 3 or 4 clusters. Only when we allow an overlap > 0.1, we obtain optimal clusterings with $k = 5$. For each parameter setting, we show the *mean*(w) of the best clustering in Figure 8.3 and Figure 8.4.

Figure 8.3 shows results for disjoint clusterings (*maxOverlap* $= 0$). As expected, we find better solutions when we relax the completeness constraint by lowering the *minCompl* parameter. Also allowing smaller sets (a lower *minSupport*) improves the quality of the solutions.

The effect of allowing overlaps can be seen in Figure 8.4. As in the disjoint setting, the quality of the solution improves with lower *minCompl*. Increasing *maxOverlap* leads to an additional improvement.

Overall, our experiments show that the quality of the resulting clustering can be increased by relaxing the completeness constraint, allowing overlapping clusters, and allowing smaller base clusters.

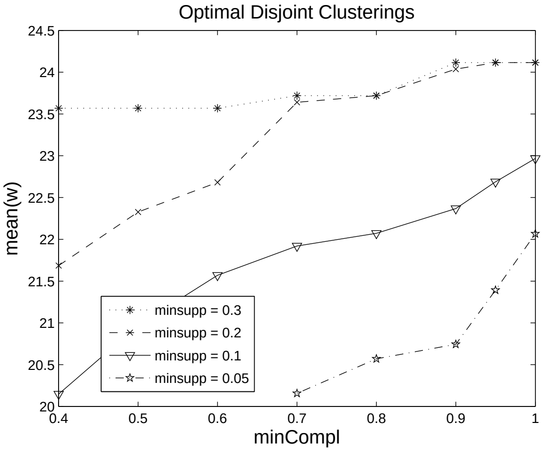


Figure 8.3: Mean within-cluster distance (y-axis) for varying *minCompl* (x-axis) of the optimal disjoint clusterings with varying *minSupport*.

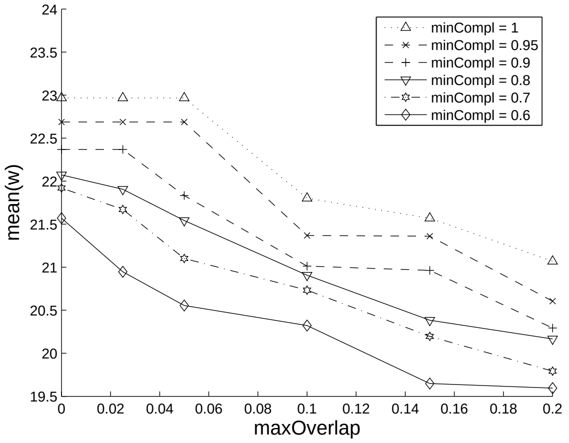


Figure 8.4: Mean within-cluster distance (y-axis) for varying *maxOverlap* (x-axis) of resulting itemsets (clusters) with a *minSupport* of 0.1.

8.4.1 Scalability

In a second batch of experiments, we focus on the scalability of our approach. This is known to be a big challenge for integer linear programs. We experiment on two publicly available datasets on thyroid disease[6] ($m =$ 2,659 examples, 16 categorical and 6 numerical attributes[7]) and on forest cover type[8] ($m =$ 10,000 examples, 44 binary and 10

[6] http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/

[7] We discretized the numerical attribute age into five (equally sized) values. Also, we removed examples with no numerical attribute values.

[8] http://archive.ics.uci.edu/ml/datasets/Covertype

numerical attributes). The pairwise Euclidian distances are computed on the normalized numerical attributes. On the thyroid dataset, we determine n=12,134 frequent itemsets with a *minSupport* of 0.1 on the categorical attributes. We did experiments on the entire thyroid dataset and two subsets: one with m=500 examples and n=12,134 base clusters, and the other one with m=1,000 examples and again n=12,134 base clusters. On the cover type dataset, we determine n=709 frequent itemsets with a *minSupport* of 0.05 on the categorical attributes.

As in the previous experiments, we tested different values for k, *minCompl*, and *maxOverlap*. For all tested parameter settings on the dementia data set, time varies from 0.4 seconds to almost 15 hours per problem instance, with 10 minutes on average and 31.7 seconds for the median.[9] On the thyroid dataset (see Table 8.6), the majority (86.8%) of problems is solved in less than 500 seconds, 78% in less than 300 seconds, and 36% in less than 60 seconds (minimum is 7.4 seconds). We interrupt runs that take longer than six hours and mark them with an ×.

The following observations from more than 500 runs on the benchmark (thyroid) and the real-world dataset (dementia) shed some light on the behavior of the optimizer depending on some key parameters:

- The running times appear to scale roughly linearly in the number of base clusters n (see Figure 8.5) and the number of examples m (see rows in Table 8.6).

- Relaxing the *maxOverlap* constraint from zero (disjoint clusters) to slightly larger values typically leads to an (often sharp) increase in the running times. However, they may decrease again for larger values (compare Figure 8.5 and 8.6).

- A similar observation can be made for *minCompl*: Relaxing the completeness requirement from one to slightly smaller values typically leads to an (often sharp) increase in the running times. However, further reducing *minCompl* often does not change the running times too much (compare Figures 8.5 and 8.7). Setting *minCompl* $= 1$ is easier to solve because the problem solver has to search only in the solution space where $y_i = 1$ for all i (compare (iv) in Table 8.3). *minCompl* typically harms performance

- On the thyroid dataset and the cover type dataset, allowing overlaps in combination with a lower *minCompl* increases the runtime dramatically (compare results for *minCompl* $= 0.9$ and *maxOverlap* $= 0.05$). For those settings no optimal solution could be obtained in reasonable time.

- The behavior in terms of k (between *minK* and *maxK*) is highly non-monotonic: A problem instance may be extremely hard for a certain k, whereas it may become easy again for $k + 1$. This may be explained by the "puzzle" that has to be solved: It may be impossible to reach a certain required completeness for a smaller number of larger "tiles". However, given a larger number of smaller "tiles", it may become possible again. The precise behavior is clearly dependent on the available base clusters.

[9] On a 512 MB RAM (900 MHz) machine

			Thyroid Dataset			Covertype
Parameters			number of examples			no. of exampl.
k	$maxO$	$minC$	500	1000	2659	$10,000$
2	0	1	8.8	15.1	70.3	0.8
4	0	1	*8.1	*13.5	*49.8	0.8
5	0	1	10.5	14.7	58.2	0.8
2	0	0.9	31.7	58.3	135.0	3.9
4	0	0.9	*105.4	*169.2	*420.0	3.7
5	0	0.9	20.8	159.0	264.9	3.6
2	0	0.8	38.6	168.9	176.4	4
4	0	0.8	356.6	116.3	547.5	4
5	0	0.8	23.7	53.9	281.7	3.7
2	0.05	1	25.1	40.4	84.6	2.1
4	0.05	1	*21.4	*38.8	*101.5	2.1
5	0.05	1	26.0	42.0	92.9	2.1
2	0.05	0.9	157.5	296.5	346.8	5368.5
4	0.05	0.9	×	×	×	×
5	0.05	0.9	119.1	14382.3	×	6441.8
2	0.1	1	21.3	40.4	107.1	2.1
4	0.1	1	*28.8	*63.3	*186.7	2.1
5	0.1	1	42.6	69.3	206.7	2.2
2	0.1	0.9	65.6	256.4	632.2	641.4
4	0.1	0.9	1351.3	1030.1	17408.9	1168.6
5	0.1	0.9	133.0	2005.0	×	×

Table 8.6: Running times (in seconds) of the thyroid dataset and the cover type dataset for varying parameters. We test on three different dataset sizes with 500, 1000 and 2659 examples. $*$ indicates that no solution exists for these parameters. $\times$ indicates that the experiment was stopped after six hours.

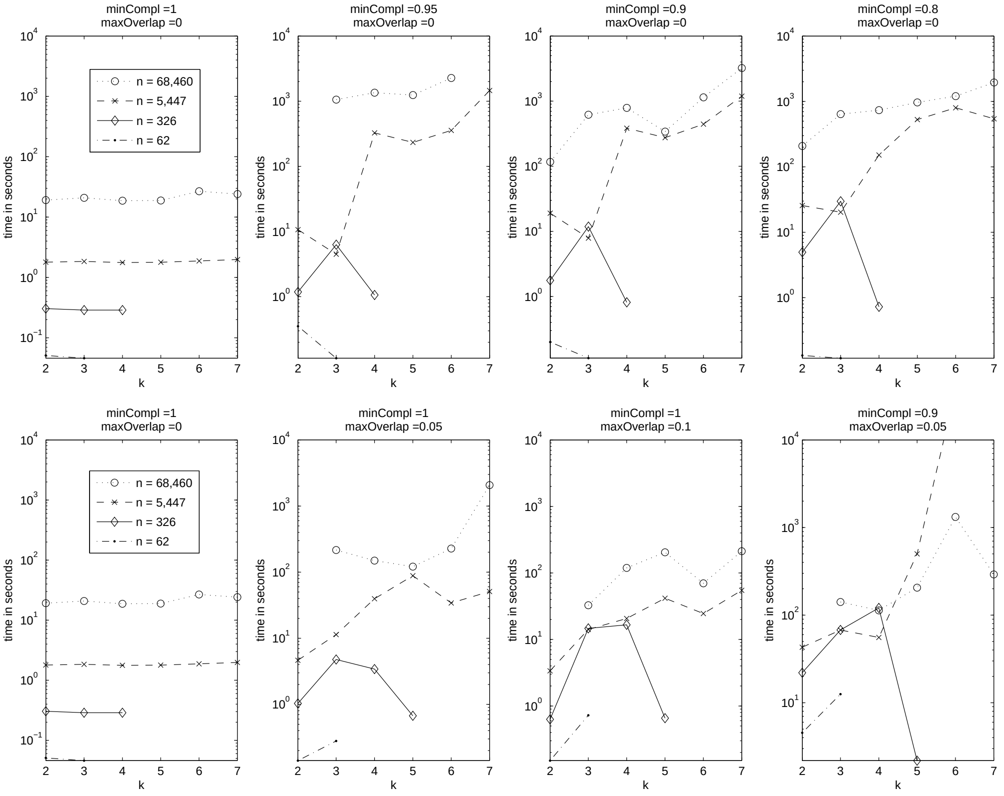


Figure 8.5: Runtime (in seconds) of disjoint clustering on dementia data for different k and different numbers of base clusters n.

Experiments with set-level constraints (detailed results not shown) indicate that they either make a problem insolvable (which can be determined very fast) or do not impact running times, because they constitute only a very small fraction of the constraints. Simple set-level constraints like $C_j \in C$, however, simplify the set of base levels a priori and thus speed up the overall process.

In summary, our experiments on two datasets showed that the running times depend on the structure of the problems stronger than on their size. Although the majority of tested problem instances was computable within a reasonable time, we found some instances that were more difficult to compute than others. For the latter cases, it may be an option to set a runtime limit and output a near-optimal solution.

8.5 Related Work

Constrained clustering has been extensively investigated over the past few years. We refer to Section 3.3 for a brief introduction. Many contributions focus on incorporating background knowledge in the form of instance-level constraints (e.g., [87]). More recent work investigates set-level and other types of constraints (e.g, [30]). Davidson *et al.* [30, 31] study the computational complexity of finding a feasible solution for clustering with constraints and show that finding a feasible solution is NP-complete for a combination of instance-level and cluster-level constraints. Clustering has been approached with linear programs before [32, 77]. However, these approaches start from a number of instances they want to assign to clusters, whereas our approach starts from a set of possible base clusters. Demiriz *et al.* [32] use linear programs to solve k-means with the constraint that each cluster has to contain a minimum number of points. This approach is extensible to pairwise constraints. In their experiments they show that it is feasible to solve constrained clustering by linear programming even for large datasets.

Itemset classified clustering has been introduced by Sese *et al.* [82]. They start from a dataset with feature attributes and objective attributes. In our case, the PETs-voxels correspond to the objective attributes and the psychological data corresponds to the feature attributes. Sese *et al.* focus on 2-clusterings and maximize the interclass variance between the two groups. Our approach handles a more general setting and can also find k-clusterings with $k > 2$.

Although it may appear related at first glance, the approach is different from clustering approaches for association rules (e.g., by An *et al.* [6]) in its goal of constrained clustering (not summarization of pattern mining results). However, set covering and set packing approaches may also be useful for summarizing itemsets and association rules.

Chaudhuri *et al.* [23] mention the weighted set packing problem in the context of finding a partition consisting of valid groups maximizing a benefit function. However, they solve the problem in a greedy fashion and thus do not aim for a global optimum.

8.6 Discussion and Conclusion

We presented an approach to constrained clustering based on integer linear programming. The main assumption is that a (possibly large) set of candidate clusters is given in advance, and that the task is then to construct a clustering by selecting a suitable subset. The construction of a suitable clustering from candidate clusters can be constrained in various ways. *Clustering constraints* allow specifying the degree of completeness of a clustering, the allowed overlap of clusters, and whether encompassing clusters are acceptable (hierarchical clusterings). In contrast, *set-level constraints* let the user explicitly state logical formulae that must hold for clusterings to be valid, for instance, that two clusters are mutually exclusive or that one cluster requires another in a clustering. Set-level constraints restrict the combinations of admissible clusters without reference to the instances. The overall quality of a clustering can be optimized in various ways: by optimizing the minimum, the mean and the median of the individual clusters' qualities in a clustering.

Our provided framework is very general and flexible and hence can be adapted to the user's needs. The user may start with the default values of $minCompl = 1.0$ and $maxOverlap = 0$. In case she wants to increase the quality of the resulting clustering, she can relax the completeness constraint to a lower value of $minCompl$ and/or allow for some multiply covered examples by increasing the $maxOverlap$ parameter. Additionally, set-level constraints may be used to exclude or include certain clusters.

Given such a set of constraints, it is then possible to map a set of constraints onto a program for integer linear programming. In this sense, the presented work stands in the tradition of other approaches to constraint-based mining and inductive query languages, where the technical complexity of the task is hidden from the users and they can still freely combine mining primitives according to their interests and preferences [19, 20, 76, 72, 34, 18].

Although integer linear programming is known to be an NP hard problem, there are fast solvers available today, making use of a wide range of different solution strategies and heuristics. Generally speaking, the base set of candidate clusters still has to be relatively small (compared to the power set of instances) to keep the optimization feasible. Vice versa, the set of constraints should not be excessively large. Contrary to the intuition, however, that the running times should depend heavily on the number of instances and the number of available clusters, we found in our experiments that the scalability in these two parameters was not as critical as expected. Other parameters, like the degree of allowed overlap, showed a much greater impact on the running times.

From a more general point of view, it is clear that problem instances with excessive running times exist, and we also encountered such instances in our experiments. In future work, we plan to address this issue by near-optimal solutions, which are an option offered by many optimization packages like Xpress-Optimizer. One possible approach could be based on a user-defined time limit: If the optimal solution can be found within the time frame, it is returned and flagged as optimal. If the time limit is exceeded and a solution was found, the best solution could be returned and flagged as near-optimal. If no solution was found within the given time, the user is informed about this outcome. In this way, the system remains transparent about the quality of its solutions.

In future work, it would be interesting to see if some of the set-level constraints could be pushed into the algorithms for solving integer linear programs, or if they have to be solved, for instance, by splitting a problem into a series of related optimization problems. One of the disadvantages of the current formulation is that the cluster quality has to be defined for each cluster in advance and not by accessing the examples during the optimization process. Also, we focus on the within-cluster distances in this work, and do not take into account between-cluster distances yet. Taking into account the between-cluster distances is possible, but would require a slightly different formulation of the optimization problem. Finally, it would be interesting to investigate whether similar techniques could be used in rule learning and related problems like subgroup discovery.

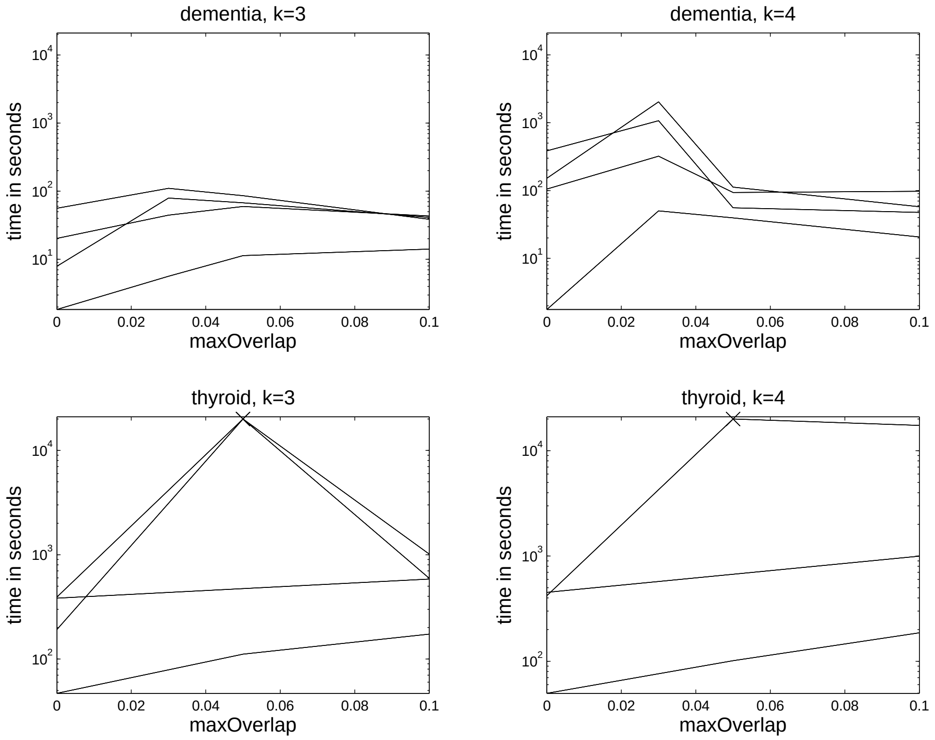


Figure 8.6: Effect of varying *maxOverlap* values on the runtime for $k = 3$ (left) and $k = 4$ (right) and $minCompl \in \{0.7, 0.8, 0.9, 1.0\}$.

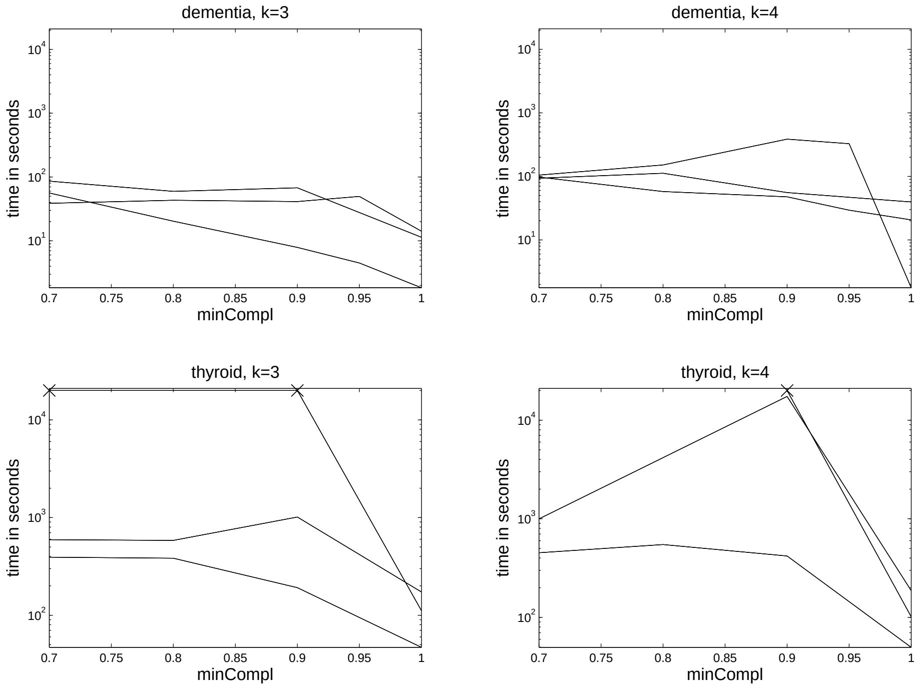


Figure 8.7: Effect of varying *minCompl* values on the runtime for $k = 3$ (left) and $k = 4$ (right) and $maxOverlap \in \{0, 0.05, 0.1\}$.

9 Conclusion

We now summarize the main contributions and give an outlook on possible directions for further research.

9.1 Summary

The main goal of this thesis was to investigate how a comprehensive analysis of medical datasets can improve the process of diagnosis. For this purpose, we analyzed two medical datasets, one on breast cancer diagnosis and one on dementia patients. We focused on two clinical tasks, the selection of an optimal test for a patient that yields an accurate diagnosis, and the correlation of image and non-image data.

For the former task, we introduced a data-efficient approach to test selection that is based on the concept of conditional mutual information. The computation of conditional mutual information relies on an accurate estimation of the underlying probability distributions. However, this estimation is often difficult to obtain, due to small datasets and the underlying computational complexity. We therefore had to find alternative ways to estimate the underlying probabilities. We described and compared a basic approach consisting of averaging mutual information estimates conditioned on individual observations and another approach where it is possible to condition on all observations at once. This is achieved by making some conditional independence assumptions represented in a Bayes net. We showed that these alternatives allow an accurate estimation of the probabilities and thus an appropriate calculation of the conditional mutual information. This then can be applied to determine which test to conduct for a patient at a certain point in time.

Many questions in medical research are related to the discovery of statistically interesting subgroups of patients. However, subgroup discovery with a single target variable is rarely sufficient in practice. Rather, more complex variants, e.g., handling costs and complex output in the form of images, are required. In this thesis, we considered such variants of subgroup discovery in the context of the clinical task of diagnosis: For breast cancer diagnosis, the task is to detect subgroups for which single modalities should be given priority over others, as indicated by a quality function. For Alzheimer diagnosis, the task is to find and analyze subgroups sharing similar PET scans. Here, the subgroup descriptions are based on features from psychological tests and background information about the patient.

For breast cancer diagnosis, we designed an algorithm that handles qualities (or costs) and finds the best subgroups of a population. By limiting the output size to the best t subgroups, it was possible to prune the search space considerably, especially for lower values of the minimum frequency (i.e., support) parameter. Therefore, the proposed algorithm SD4TS clearly outperforms the baseline algorithm used for comparison (APRIORI_SD).

On a set of Alzheimer's patients with PET scans we were able to show that subgroup discovery also works well for complex outputs such as images. We developed a workflow that first clusters images, and then applies subgroup mining to find subgroups with unusual statistical patterns in certain clusters. We also developed an alternative workflow that first determines frequent itemsets in the non-image data obtaining groups of patients that share the same test result or demographical information. In a second step, these groups are evaluated by computing the similarity of their brain scans. Thus, these groups can be considered as base clusters that can be combined into an overall clustering. We proposed a method for constrained clustering that can perform this combination while satisfying several user-defined constraints. This methods is based on translating the constraints into an integer linear program that then can be solved with standard optimization tools. Both approaches could provide useful hints for physicians and perhaps even suggest new hypotheses.

Summing up, the main contributions of the thesis are:

- a new method for test selection based on information maximization,
- new methods for subgroup discovery with costs and complex output,
- two alternative workflows for the correlation of image and non-image data,
- a new method for constrained clustering that offers a translation of a variety of user-defined constraints into an interger-linear program which then can be solved with standard optimization tools, and
- an implementation and experimentally evaluation of those methods.

Overall, we showed that subgroup discovery can be a valuable tool for the analysis of breast cancer and Alzheimer diagnosis data. Similar techniques should be applicable successfully in many other medical domains as well.

9.2 Challenges and Limitations

The main problems and limitations in the context of the HUP study were caused by the small sample size (138 examples, i.e., lesions) and the equivocality of optimal test selection. In other words, for many cases, two or more tests perform equally well in practice. On the Alzheimer data, the results of subgroup discovery depend critically on the quality of the clusterings. As clustering of PET scans is largely uncharted territory, we believe that there should be ample room for improvement in this area.

Our results show that it is promising to continue this kind of analysis in other medical fields. However, a major limitation is acquiring the necessary data. It seems that most hospitals are yet not prepared. Even though hospitals store a huge volume of data, this data is mostly not accessible for research. There are many reasons for this. First, there is a high justifiable concern about privacy in order to protect the patients. In general, the patients have to agree that their data is used for research purposes. This is no problem for new patients. However, to get agreements of past patients retrospectively is almost impossible. This can be circumvented by a de-identification of the data which is often difficult, time consuming, and expensive.

A second reason is that most of the data is not digitally accessible. Although many hospitals maintain a hospital information system (HIS) to store and administer patient records, test results, and the billing system, this HIS is not generally available for research purposes (see privacy concerns). Furthermore, this system is rarely a complete source of information. For instance, in the dementia study, most test records were only available on paper, which are archived in a huge collection of ring binders. Using this kind of data always requires manual data entry and revision.

A huge challenge is the inclusion of images. Images are owned by the patients, here an agreement of the patient is also needed. Beyond that, only few hospitals store all images in a PACS system (Picture archiving and communication system), a digital medical image management, communication, review and distribution system. Many pictures are stored only on CD-ROMs, or in the case of X-rays, only on film. There is a need for research image databases, which also store annotations of the images and provide a link to other information of the patient.

A further reason that complicates the data access, is that most patient records contain only very little structured information. Most information is free text (often with typos), whose analysis requires sophisticated text mining and natural language processing techniques. The amount of information provided in the records varies from patient to patient and strongly depends on the time and accuracy of the author.

In order to get universal results, an ultimate goal would be to combine datasets across hospitals. However, this seems to be a distant prospect, because right now it is often already very difficult to combine datasets of different departments of the same hospital.

Overall, these challenges should not sound too daunting. Our experience shows that there are already departments that are willing to invest in proper data storage and are also open to this type of analysis. Furthermore, much research in the area of health care business addresses the introduction of research databases that store patient records and images and enable their efficient management. Therefore, we expect to find better conditions in the near future.

9.3 Directions for Future Work

In future work, research could be extended towards interweaving the two methods of clustering and subgroup discovery. For instance, during the process of subgroup generation, the similarity of images could be computed “on the fly” and used for pruning. This could yield a form of semi-constrained clustering, where the base clusters do not have to be restricted to the *cover* of an itemset, which is the set of examples that exactly match the itemset description. Moreover, the base clusters can represent “itemset and friends", which are sets of examples that have approximately the same features. This can be achieved by first determining frequent itemsets and their corresponding covers. In a second step it is possible to modify the covers by either adding or removing some examples. For this task different heuristics are conceivable: a new example should only be added if part of the itemset description is matching, or in case the adding will improve the pairwise similarity of images above a certain threshold. In this context, a further idea is to work with disjoint itemset descriptions which enables the merging of two complete covers of two different itemsets.

Our approaches to test selection can be extended to a decision support system that provides the health care worker with information on which test(s) to conduct on a patient. Further ideas for the analysis of the HUP data are outlined in Section 2.1.4. A very challenging and interesting task is to design a similarity function between two patients. A good similarity function enables a new patient to be matched with similar patients, who were treated previously. Knowing their diagnosis and therapy outcome could help to diagnose new patients more accurately and advise an optimal therapy.

A different direction of research is to address the challenges discussed in the previous section. For instance, one can develop research databases that are accessible to health care workers and researchers. One important aspect of these databases is that they have to be easy to use for the health care workers, because the basis of a complete and up-to-date database is the prompt data entry of new patients. A further step is to provide a tool that integrates the database and software applications that are already used in the clinical business, *e.g.,* image viewers or image processing tools, or data mining tools.

Acknowledgements

I am indebted to a number of people. Without their support and advice, this work would not have been possible.

First of all, I want to thank Stefan Kramer, my advisor. His knowledge, support, advice, guidance and patience have helped me develop and refine my skills and knowledge in order to complete this thesis.

Many thanks Rómer Rosales, Sriram Krishnan, Bharat Rao from Siemens Medical Solutions USA. Being a part of their research group for ten months as well as access to their scientific and industrial research insight was inspiring and beneficial. Furthermore, I want to thank Siemens Medical for their generous financial support as well as Nassir Navab and Stefan Kramer for initiating this collaboration.

I want to thank my "medical advisors", Mitch Schnall from Hospital of The University of Pennsylvania in Philadelphia, Ulrike Schwarz-Boeger from Klinikum Rechts der Isar München, and Gudrun Schimpfle and Silvia Heywang-Koebrunner from Mammazentrum München for giving me insights into the diagnosis of breast cancer and discussing our results, as well as Robert Perneczky, Alexander Kurz, and Alexander Drzezga from Klinikum Rechts der Isar München for giving me insights in the diagnosis of Alzheimer's disease and the interpretation of experiment results.

Additional thanks goes out to Gudrun Zahlmann and Markus Schmitt from Siemens Healthcare Erlangen for including me in two medical trials and also for their industrial insights and financial support.

Also a big shout out to Barbara Wilhelm and Michael Ritter for introducing me to the world of linear optimization and their generous help and patience.

Moreover, I want to thank my students, whose Diploma theses, Bachelor theses and system development projects helped me in my work. In particular, I want to thank Andreas Hapfelmeier and Jana Schmidt for their great work and well functioning implementations of the Alzheimer's project.

All of my colleagues in our research group at Technische Universität München created a very enjoyable working atmosphere, and definitely increased my motivation to come to Garching. In addition, I appreciate and thank the K@ffchen-group for countless energy boosting conferences and also their senior advices about how to finish a thesis.

Thanks also to the MVV for finally extending the subway line to Garching Forschungszentrum saving me from being compressed into the smelly buses and bringing me to work earlier allowing me to accomplish even more!

Last but not least, I thank my friends and family for standing my moods, and constantly motivating and supporting me during all these years. Thanks boys and girls!

List of Figures

List of Tables

Bibliography

[1] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[3] Alzheimers‘s Disease International. The prevalence of dementia worldwide. `http://www.alz.co.uk/adi/pdf/prevalence.pdf`, 2008.

[4] American Cancer Society. Detailed guide: Breast cancer can breast cancer be found early? `http://www.cancer.org/docroot/cri/content/cri_2_4_3x_can_breast_cancer_be_found_early_5.asp`, September 2009. [Online; accessed 20-February-2010].

[5] American College of Radiology. *BI-RADS Breast Imaging Reporting and Data System, Breast Imaging Atlas*. American College of Radiology, 4th edition, 2003.

[6] Aijun An, Shakil Khan, and Xiangji Huang. Objective and subjective algorithms for grouping association rules. In *Third IEEE International Conference on Data Mining (ICDM 2003)*, pages 477– 480, 2003.

[7] Steen Andreassen. Planning of therapy and tests in causal probabilistic networks. *Artifical Intelligence in Medicine*, 4:227 – 241, 1992.

[8] Martin Atzmüller and Florian Lemmerich. Fast subgroup discovery for continuous target concepts. In Jan Rauch, Zbigniew W. Ras, Petr Berka, and Tapio Elomaa, editors, *Proceedings of 18th International Symposium Foundations of Intelligent Systems (ISMIS 2009)*, volume 5722 of *Lecture Notes in Computer Science*, pages 35–44. Springer, 2009.

[9] Martin Atzmüller and Frank Puppe. SD-Map - a fast algorithm for exhaustive subgroup discovery. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Proceedings of the 10th European Conference on Principles and Practice in Knowledge Discovery (PKDD 2006)*, volume 4213 of *Lecture Notes in Computer Science*, pages 6–17. Springer, 2006.

[10] Martin Atzmüller and Frank Puppe. Semi-automatic refinement and assessment of subgroup patterns. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS 2008)*, pages 323–328, 2008.

[11] Arindam Banerjee and Joydeep Ghosh. Clustering with balancing constraints. In *Constrained Clustering: Algorithms, Applications and Theory*. Chapman & Hall/CRC Press, 2008.

[12] Heinrich Bär and Juergen Sauer. Diagnostik und Therapie häufiger Demenzen. In *Ärzteblatt Thüringen*, pages 413–414, Jena, Germany, 2006.

[13] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Algorithms, Applications and Theory*. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC Pres, 2008.

[14] Joachim Baumeister, Martin Atzmüller, and Frank Puppe. Inductive learning for case-based diagnosis with multiple faults. In Susan Craw and Alun D. Preece, editors, *Advances in Case-Based Reasoning, Proceedings of the 6th European Conference (ECCBR 2002)*, volume 2416 of *Lecture Notes in Computer Science*, pages 28–42. Springer, 2002.

[15] Pavel Berkhin. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg, 2006.

[16] Christian Böhm, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density connected clustering with local subspace preferences. In *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM 2004)*, pages 27–34, Washington, DC, USA, 2004. IEEE Computer Society.

[17] Mario Boley and Henrik Grosskreutz. Non-redundant subgroup discovery using a closure system. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009)*, volume 1, pages 179–194. Springer, 2009.

[18] Francesco Bonchi, Fosca Giannotti, and Dino Pedreschi. A relational query primitive for constraint-based pattern mining. In *Constraint-Based Mining and Inductive Databases*, pages 14–37. Springer, 2004.

[19] Jean-Francois Boulicaut and Baptiste Jeudy. Constraint-based data mining. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 399–416. Springer, 2005.

[20] Jean-Francois Boulicaut and Cyrille Masson. Data mining query languages. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 715–727. Springer, 2005.

[21] M J Chandler, L H Lacritz, L S Hynan, H D Barnard, G Allen, M Deschner, M F Weiner, and C M Cullum. A total score for the CERAD neuropsychological battery. *Neurology*, 65(1):102–6, 2005.

[22] Oliver Chapelle, Bernhard Schölkopf, and A. Zien. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006.

[23] Surajit Chaudhuri, Anish Das Sarma, Venkatesh Ganti, and Raghav Kaushik. Leveraging aggregate constraints for deduplication. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, pages 437–448, 2007.

[24] Jake Yue Chen, Chang Yu Shen, and Andrey Y. Sivachenko. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pacific Symposiun on Biocomputing*, 11:367–378, 2006.

[25] Peter Clark and Tim Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

[26] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[27] Giorgio Corani, Chris Edgar, Isabelle Marshall, Keith Wesnes, and Marco Zaffalon. Classification of dementia types from cognitive profiles data. In *Proc. of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pages 470–477. Springer, 2006.

[28] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Interscience, 1991.

[29] Dash Optimization. Xpress-MP, 2009.

[30] Ian Davidson and S. S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. In *Proceedings of the Fifth SIAM International Conference on Data Mining (SDM 2005)*, pages 138–149, 2005.

[31] Ian Davidson and S. S. Ravi. The complexity of non-hierarchical clustering with instance and cluster level constraints. *Data Mining and Knowledge Discovery*, 14(1):25–61, 2007.

[32] Ayhan Demiriz, Kristin Bennett, and Paul S. Bradley. Using assignment constraints to avoid empty clusters in k-means clustering. In *Constrained Clustering: Algorithms, Applications and Theory*, pages 201–220. Chapman & Hall/CRC Press, 2008.

[33] Peter Doubilet. A mathematical approach to interpretation and selection of diagnostic tests. *Medical Decision Making*, 3:177 – 195, 1983.

[34] Saso Dzeroski, Ljupco Todorovski, and Peter Ljubic. Inductive queries on polynomial equations. In Jean-François Boulicaut, Luc De Raedt, and Heikki Mannila, editors, *Constraint-Based Mining and Inductive Databases*, pages 127–154. Springer, 2004.

[35] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, pages 226–231. AAAI Press, 1996.

[36] Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.

[37] Glenn Fung and Jonathan Stoeckel. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. *Knowledge and Information Systems*, 11(2):243–258, 2006.

[38] Michael R. Garey and David S. Johnson. *Computers and Intractability.* Freeman, 1979.

[39] Corrado Gini. Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955), 1912.

[40] Paul Glasziou and Jørgen Hilden. Test selection measures. *Medical Decision Making*, 9:133–141, 1989.

[41] Peter C Gøtzsche and Margrethe Nielsen. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, 4, 2009.

[42] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008)*, volume 1, pages 440–456, 2008.

[43] Henrik Grosskreutz and Stefan Rüping. On subgroup discovery in numerical domains. *Data Mining Knowledge Discovery*, pages 210–226, 2009.

[44] Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for optimal decision making. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, pages 149–154, 2007.

[45] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of ACM SIGMOD Intl. Conference on Management of Data*, pages 1–12, New York, NY, USA, 2000. ACM.

[46] Andreas Hapfelmeier, Jana Schmidt, Marianne Mueller, Robert Perneczky, Alexander Kurz, Alexander Drzezga, and Stefan Kramer. Interpreting PET scans by structured patient data: A data mining case study in dementia research. In *Eighth IEEE International Conference on Data Mining 2008 (ICDM 2008)*, pages 213–222, 2008.

[47] Sylvia Heywang-Köbrunner, David D. Dershaw, and Ingrid Schreer. *Diagnostic Breast Imaging.* Georg Thieme Verlag, 2000.

[48] C. P. Hughes, L. Berg, and W. L Danzinger. A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, 140:566–572, 1982.

[49] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Clustering data: A review. In *ACM Computing Surveys*, volume 31, pages 264–323, 1999.

[50] Karsten Juhl Jørgensen and Peter C Gøtzsche. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *British Medical Journey (BMJ)*, 339(b2587), July 2009.

[51] John G. Kalbfleisch. *Probability and Statistical Inference: Vol. 2: Statistical Inference*. Springer, 1985.

[52] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data. An introduction to cluster analysis*. Wiley, 1990.

[53] Branko Kavšek and Nada Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[54] Willi Klösgen. Explora: a multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pages 249–271, Menlo Park, CA, USA, 1996. American Association for Artificial Intelligence.

[55] Kooperationsgemeinschaft Mammographie des Deutschen Krebsforschungszentrums, Krebsinformationsdienst, and Deutsche Krebshilfe. Mammographie-Screening. Früherkennung von Brustkrebs. Was Sie darüber wissen sollten., 2009.

[56] Nada Lavrač, Branko Kavsek, Peter A. Flach, and Ljupco Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[57] Nada Lavrač, Filip Železný, and Peter A. Flach. RSD: Relational subgroup discovery through first-order feature construction. In S. Matwin and C. Sammut, editors, *Proceedings of the 12th International Conference on Inductive Logic Programming*, volume 2583 of *Lecture Notes in Artificial Intelligence*, pages 149–165. Springer, 2003.

[58] Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Data 2008 Part II (ECML PKDD 2008)*, volume 5212 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2008.

[59] Dennis Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[60] A. Lobo, L.J. Launer, L. Fratiglioni, K. Andersen, A. Di Carlo, and M.M. Breteler. Prevalence of dementia and major subtypes in europe. a collaborative study of population-based cohorts. *Neurology*, 54(11 Suppl 5):4–9, 2000.

[61] David MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[62] Subraman Mani, W.R. Shankle, Michael J. Pazzani, Padhraic Smyth, and Malcolm B. Dick. Differential diagnosis of dementia: A knowledge discovery and data mining (KDD) approach. *American Medical Informatics Association (AMIA) Annual Fall Symposium*, 1997.

[63] Vasileios Megalooikonomou, James Ford, Li Shen, Fillia Makedon, and Andrew Saykin. Data mining in brain imaging. *Statistical Methods in Medical Research*, 9:359–394, 2000.

[64] Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM 2004)*, pages 483–486, 2004.

[65] Randolph A. Miller. Medical diagnostic decision support systems—past, present, and future. *Journal of the American Medical Informatics Association*, 1(1):8–27, 1994.

[66] John C. Morris. CDR: Current version and scoring rules. *Neurology*, 43:2412–2414, 1993.

[67] John C. Morris, R.C. Mohs, and H. Rogers. CERAD: clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacol Bull*, 4:641–652, 1988.

[68] Marianne Mueller and Stefan Kramer. Integer linear programming models for constrained clustering. In Geoffrey Holmes Bernhard Pfahringer and Achim Hoffman, editors, *Proceedings of the 13th International Discovery Science (DS 2010)*, pages 159–173, 2010.

[69] Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. Data-efficient information-theoretic test selection. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME 2009)*, pages 410–415, 2009.

[70] Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. Subgroup discovery for test selection: A novel approach and its application to breast cancer diagnosis. In *Advances in Intelligent Data Analysis VIII*, volume 5651, pages 119–130, 2009.

[71] National Institute of Aging. Alzheimer's disease: Unraveling the mystery. `http://www.nia.nih.gov/Alzheimers/Resources/HighRes.htm`, February 2009. [Online; accessed 20-February-2010].

[72] Siegfried Nijssen and Luc De Raedt. IQL: A proposal for an inductive query language. In *Knowledge Discovery in Inductive Databases*, volume 4747, pages 189–207. Springer, 2007.

[73] Petra Kralj Novak and Nada Lavrač anf Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of constrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

[74] Carlos Ordonez, Norberto Ezquerra, and Cesar A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):259–289, 2006.

[75] Robert Perneczky, A. Drzezga, J Diehl-Schmid, G Schmid, A Wohlschlager, S Kars, T Grimmer, S Wagenpfeil, A Monsch, and A Kurz. Schooling mediates brain reserve in Alzheimer's disease: findings of FDG PET. *Journal of Neurology, Neurosurgery and Psychiatry (JNNP)*, 77:1060–1063, 2006.

[76] Luc De Raedt. A perspective on inductive databases. *SIGKDD Explorations*, 4(2):66–77, 2002.

[77] Burcu Saglam, F. Sibel Salman, Serpil Sayin, and Metin Turkay. A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research*, 173(3):866–879, 2006.

[78] Jana Schmidt, Andreas Hapfelmeier, Marianne Mueller, Robert Perneczky, Alexander Kurz, Alexander Drzezga, and Stefan Kramer. Interpreting PET scans by structured patient data: A data mining case study in dementia research. *Journal of Knowledge and Information Systems (KAIS)*, 24(1):149–170, 2010.

[79] Alexander Schrijver. *Theory of Linear and Integer Programming.* John Wiley & Sons, 1998.

[80] Danielle Sent and Linda G. van der Gaag. On the behaviour of information measures for test selection. In *Proceedings of 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, pages 316–325, 2007.

[81] Jun Sese, Yukinori Kurokawa, Morito Monden, Kikuya Kato, and Shinichi Morishita. Constrained clusters of gene expression profiles with pathological features. *Bioinformatics*, 20(17):3137–3145, 2004.

[82] Jun Sese and Shinichi Morishita. Itemset classified clustering. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004)*, volume 3202 of *Lecture Notes in Computer Science*, pages 398–409. Springer, 2004.

[83] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[84] Michael Steinbach, George Karypis, and Bipin Kumar. A comparison of document clustering techniques. Technical report, University of Minnesota, May 2000.

[85] Filip Železný. *RSD – a system for relational subgroup discovery through first-order feature construction – user‘s manual*, 03 2003. version 1.

[86] Filip Železný and Nada Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1-2):33–63, 2006.

[87] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 577–584, 2001.

[88] P. Roy Walker, Brandon Smith, Qing Y. Liu, Fazel Famili, Julio Valdes, Ziying Liu, and Boleslaw Lach. Data mining of gene expression changes in Alzheimer brain. *Artificial Intelligence in Medicine*, 31(2):137–154, 2004.

[89] Ian Watson and Farhi Marir. Case-based reasoning: A review. *The Knowledge Engineering Review*, 9:327–354, 1994.

[90] World Health Organization. *ICD-10 : International Statistical Classification of Diseases and Related Health Problems (Tenth Revision)*. World Health Organization, Geneva, Switzerland, 2 edition, 2005.

[91] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In Henryk Jan Komorowski and Jan M. Zytkow, editors, *PKDD*, volume 1263 of *Lecture Notes in Computer Science*, pages 78–87. Springer, 1997.

[92] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2007.

[93] Alice X. Zheng, Irina Rish, and Alina Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. In *Proceedings of UAI-05, 21st Conference on Uncertainty in Artificial Intelligence*, pages 675–682, 2005.

[94] Zhiqiang H. Zheng and Balaji Padmanabhan. On active learning from data acquisition. In *Proceedings of ICDM-2002, Second IEEE International Conference on Data Mining*, pages 562–569, 2002.

[95] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin – Madison, 2007.

[96] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*, volume 3 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2009.