TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

Data Mining and Machine Learning Methods for High-dimensional Patient Data in Dementia Research: Voxel Features Mining, Subgroup Discovery and Multi-view Learning

Rui Li

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

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Nils Thuerey

Prüfer der Dissertation:

- 1. Univ.-Prof. Dr. Burkhard Rost
- 2. Univ.-Prof. Dr. Stefan Kramer Johannes Gutenberg-Universität Mainz

Die Dissertation wurde am 22.06.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 01.11.2016 angenommen.

This page is intentionally left blank.

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and ${\rm discussions.}$ München, den 03.05.2016

The scientist is motivated primarily by curiosity and a desire for truth.

Irving Langmuir

Abstract

Today, patient data often includes large amounts of structured information, such as neuroimaging data, neuropsychological test results, demographic variables, etc. Human beings, however, cannot analyze so much information, at least not without the help of modern data mining and machine learning methods. Given the diverse sources of information, computerized methods show a great promise to help clinicians to discover hidden patterns in disease data. Thus, computer-aided diagnosis (CAD) is a tool to assist clinicians in taking diagnosis decisions. The goal of the thesis is to devise computer algorithms that can uncover useful patterns, which can further be used to build predictive models.

Positron emission tomography (PET), one type of neuroimaging data, is of great importance in the diagnosis of dementia. One specific application of PET scans is to differentiate between different types of dementia, such as mild cognitive impairment (MCI) against Alzheimer's disease (AD). We develop a Gaussian mixture model with a model selection approach to automatically classify them based on some training samples. In a comparative study, the results show that the proposed method outperforms two baseline methods. A different use of PET scans is to predict the likelihood of dramatic cognitive decline, i.e., the progression from MCI to full-blown AD. To this end, a combined analysis of survival analysis together with an infinite Gaussian mixture model is presented to find discriminative brain voxels that are related to cognitive degradation. The experiments indicate that the proposed algorithm can extract critical brain regions associated with dementia progression. Additionally, non-imaging variables are brought in to complement the decision making. To make full use of imaging and non-imaging variables, a multi-view stacking approach is suggested to learn a hierarchical classification model on two levels, i.e., a base level and a meta level. The experimental results demonstrate that various types of data, also called views, can be beneficial if the correlation of views at the base and the meta level is within a certain range.

To uncover interesting rules hidden in data, we develop two new subgroup discovery algorithms. Although subgroup discovery has been studied within the area of data mining for the last twenty years, some open problems still remain. One of these problems is the huge search space for hypothesis exploration, which can be infeasible

to be explored. Another problem is the redundancy contained within the resulting rule set. To this end, an optimization based method and a topic modeling based method are devised to efficiently find rules defining subgroups with low redundancy.

Throughout the thesis, we propose algorithms for analyzing PET scans, multiview stacking for learning patterns in data from diverse information sources and subgroup discovery algorithms to find interesting rules. These algorithms provide us with new tools to gain knowledge based on data of rich and complex structure.

Zusammenfassung

Heutzutage enthalten Patientendaten häufig eine große Menge von strukturierten Informationen, wie beispielsweise Neuroimaging-Daten, neuropsychologische Testergebnisse, demographische Variablen, usw. Allerdings können Menschen so viele Informationen nicht ohne die Hilfe von modernem Datamining und maschinellen Lernverfahren analysieren. Mit Hilfe solch reichhaltiger und vielfältiger Informationsquellen können Computerwissenschaftler computergestützte Methoden entwickeln, um verborgenes Wissen und Muster zu entdecken, welche Ärzten dabei helfen können, tiefe Einblicke in Krankheiten und Krankheitsverlaufe zu gewinnen. Deswegen ist Computer-Aided Diagnosis (CAD) ein Werkzeug, um Ärzten bei schwierigen Diagnose entscheidungen zu helfen. Daher ist das Ziel dieser Doktorarbeit, Algorithmen zu entwerfen, die nützliche Muster entdecken, welche für die Erstellung von Modellen für die Vorhersage von zukünftigen Ereignissen verwendet werden können.

Positronen Emissions Tomographie (PET), eine Art von Bildgebung, ist von großer Bedeutung bei der Diagnose von Demenz. Eine Anwendung ist, zwischen verschiedenen Typen von Demenz, wie milder kognitiver Beeinträchtigung (MCI) und der Alzheimer Krankheit (AD), zu unterscheiden. Wir entwickeln ein Gaussian Mixture Model mit einem Model Selection, um eine automatisierte Klassifizierung zu ermöglichen, basierend auf einigen Trainingsbeispielen. Das Ergebnis zeigt, dass das vorgeschlagene Verfahren die damit verglichenen zwei Baseline-Methoden übertrifft. Die andere Anwendung von PET ist, die Wahrscheinlichkeit für einen dramatischen Rückgang der kognitiven Leistungsfähigkeit in naher Zukunft, also die Progression von MCI zu vollentwickelter AD, vorherzusagen. Zu diesem Zweck wird eine Kombination aus Überlebensanalyse und einem Infinite Gaussian Mixture Model prasentiert, um unterschiedliche Gehirn-voxel zu finden, die mit dem kognitiven Abbau zusammenhangen. Das Ergebnis zeigt, dass der vorgeschlagene Algorithmus einige kritische Hirnregionen extrahieren kann, die beim Fortschreiten von Demenz relevant sind. Des Weiteren bringen die nicht-bildgebenden Variablen ebenfalls ergänzende Informationen, um die Entscheidungsfindung zu begünstigen. Um die bildgebenden und nicht-bildgebenden Variablen besser auszunutzen, wird ein Multi-View-Stacking Ansatz empfohlen, um ein Klassifizierungsmodell auf zwei Ebenen, d.h., Basisebene und Meta-Ebene, zu entwerfen. Die Ergebnisse zeigen, dass verschiedene Arten von

Daten (Views) von Vorteil sein können, wenn die Korrelation zwischen Views der Basis- und Meta-Ebene in einem bestimmten Bereich liegt.

Um einige interessante Regeln aufzudecken, die in den Daten verborgen sind, entwickeln wir zwei neue Subgroup Discovery Algorithmen. Obwohl der Bereich Subgroup Discovery im Datamining schon rund zwanzig Jahre existiert, gibt es immer noch einige offene Probleme. Ein Problem ist der riesige Suchraum, der untersucht werden muss, was bei einem relativ komplexen unmöglich werden kann. Das andere Problem ist die Redundanz, die im resultierenden Regelsatz enthalten ist. Deswegen wurde ein Optimierungs-basiertes Verfahren und ein Topic Model basiertes Verfahren entwickelt, um Subgroup Discovery Regeln mit niedriger Redundanz effizient herauszufinden.

Im Rahmen der Doktorarbeit stellen wir Algorithmen für die Analyse von *PET Scans*, *Multi-View Stacking* für das Untersuchen vielfältiger Informationsquellen und *Subgroup Discovery* Ansätze für das Hervorbringen interessanter Regeln vor. Sie liefern uns ein Werkzeug, um Wissen aus strukturierten Daten zu gewinnen.

Acknowledgments

I would like to thank Prof. Dr. Stefan Kramer for his support and guidance over the past years. I have benefitted tremendously from his broad knowledge in several fields. He always encouraged me to stay curious and open-minded in academia.

Special thanks also go to Prof. Dr. Robert Perneczky who is another excellent advisor. I would have great difficulty in understanding some knowledge in medicine without his help. During my stay at Imperial College London, he spent a substantial amount of time with me on correcting a paper and discussing experimental results.

I further thank Prof. Dr. Alexander Drzezga who had given me many valuable suggestions in medicine. I would not be able to comprehend nuclear imaging technique well without his support. I would like to thank the colleagues from Technische Universität München (TUM) School of Medicine, such as Prof. Dr. Markus Schwaiger, Prof. Dr. Alexander Kurz, Dr. Igor Yakushev and Dr. Stefan Förster. In addition, I also enjoyed the time with colleagues at TUM. Andreas Hapfelmeier and Jana Schmidt assisted me at the very beginning of my PhD study. It was a pleasure to work with other colleagues as well, such as Dr. Madeleine Seeland, Fabian Buchwald, Dr. Jörg Wicker, Dr. Tobias Girschick, Dr. Lothar Richter, Matthias Böck, Constanze Schmitt and Dr. Marianne Müller. The TUM Graduate School of Information Science in Health offered me very generous support throughout my PhD in many ways. During my research stay at Mainz, I came to know Dr. Andreas Karwath, Michael Geilke, Dr. Junming Shao, Zahra Ahmadi and Sophie Burkhardt. It was great fun to work with them.

I am also very grateful to Prof. Dr. Burkhard Rost to act as an examiner of my thesis work. In addition, he offered me a feature selection project in the context of membrane proteins in 2010. In this way, I got the opportunity to learn something about an important topic from bioinformatics.

Last, but not least, I am very thankful to my parents, relatives and friends, who always stood by my side all these years.

Publications

Parts of this thesis have been pre-published:

Conferences

- R. Li, A. Hapfelmeier, J. Schmidt, R. Perneczky, A. Drzezga, A. Kurz and S. Kramer. A Case Study of Stacked Multi-view Learning in Dementia Research. In: Proceedings of the 13th Conference on Artificial Intelligence in Medicine, pages 60–69, Bled, Slovenia, Lecture Notes in Computer Science (LNCS), Springer, 2011.
- R. Li and S. Kramer. Efficient Redundancy Reduced Subgroup Discovery via Quadratic Programming. In: The 15th International Conference on Discovery Science, pages 125–138, Lyon, France, Lecture Notes in Artificial Intelligence (LNAI), Springer, 2012.
- R. Li, Z. Ahmadi and S. Kramer. Constrained Latent Dirichlet Allocation for Subgroup Discovery with Topic Rules, accepted as long paper, pages 519–524, Prague, Czech Republic, ECAI: 21st European Conference on Artificial Intelligence, IOS Press, 2014.

Journals

- R. Li, R. Perneczky, A. Drzezga and S. Kramer. Efficient Redundancy Reduced Subgroup Discovery via Quadratic Programming, accepted as invited best Discovery Science 2012 papers, JIIS: Journal of Intelligent Information Systems, 44(2): 271–288, Springer, 2015.
- R. Li, R. Perneczky, I. Yakushev, S. Förster, A. Kurz, A. Drzezga and S. Kramer. Gaussian mixture models and model selection for [18F] fluorodeoxy-glucose positron emission tomography classification in Alzheimer's disease, 10(4), PLoS ONE, 2015 April 28.

Open Access Scientific Publication Repository (hosted by Cornell University): arXiv

• R. Li, R. Perneczky, A. Drzezga and S. Kramer. Survival Analysis, The Infinite Gaussian Mixture Model, FDG-PET and Non-imaging Data in The Prediction of Progression in Mild Cognitive Impairment, 2015

Contents

1	Intro	Introduction			
	1.1	Minin	g Imaging Data	1	
		1.1.1	Dementia Classification	2	
		1.1.2	Mild Cognitive Impairment Progression	2	
	1.2	Minin	g Non-imaging Data	3	
	1.3	Comb	ining Imaging and Non-imaging Data	3	
2	Background on Dementia and Alzheimer's Disease 6				
	2.1	General Introduction to Alzheimer's Disease			
	2.2	Neuropsychological Tests and Computer-aided Diagnosis 10			
	2.3	3 Data Used in Experiments			
		2.3.1	Medical Data	13	
		2.3.2	Benchmark UCI Data	13	
3	State-of-the-art in Data Mining				
	3.1	Minin	g Imaging Data	15	
		3.1.1	Introduction to Functional Imaging Techniques	15	
		3.1.2	Related Work in Mining Voxel Data	19	
		3.1.3	Model Selection in Neuroscience	21	
	3.2	3.2 Subgroup Discovery		21	
		3.2.1	Introduction to Association Rule Mining	23	
		3.2.2	Introduction to Subgroup Discovery	25	
		3.2.3	Subgroup Discovery Algorithms	27	
		3.2.4	Evaluation Measures	29	
		3.2.5	Challenges in Subgroup Discovery	30	

Contents

	3.3	Multi-	view Learning	31
		3.3.1	Multi-view Learning Related Work	31
		3.3.2	Stacked Multi-view Learning	33
4	Mini	ng Disc	riminative Imaging Features	36
	4.1	Gauss	ian Mixture Model and Model Selection: A Classification Case	
		Study		36
		4.1.1	Motivation	36
		4.1.2	Materials and Methods	37
		4.1.3	Experiments	48
		4.1.4	Discussion	56
		4.1.5	Conclusions	59
	4.2	Surviv	val Analysis: A Follow-up Case Study	59
		4.2.1	Motivation	59
		4.2.2	Materials and Methods	59
		4.2.3	Experiments	66
		4.2.4	Discussion	67
		4.2.5	Conclusions	69
5	Sub	group D	viscovery	71
	5.1	Subgr	oup Discovery via Optimization	71
		5.1.1	Motivation	71
		5.1.2	Subgroup Discovery via Quadratic Programming (SDVQP) $% \left(\frac{1}{2}\right) =\left(\frac{1}{2}\right) \left(\frac{1}{2}\right)$	72
		5.1.3	Experiments	81
		5.1.4	Conclusions	87
	5.2	Subgr	oup Discovery via Topic Modeling	89
		5.2.1	Motivation	89
		5.2.2	Introduction to Topic Modeling	90
		5.2.3	Constrained Latent Dirichlet Allocation for Subgroup Discov-	
			ery with Topic Rules	94
		5.2.4	Experiments	99
		5.2.5	Conclusions	102
6	Mult	ti-view l	Learning: Combining Imaging and Non-Imaging Data	105
	6.1	Motiv	ation	105

Contents

6.2 Stacked Multi-View Learning			d Multi-View Learning	106
	0.2	6.2.1	_	106
			Stacking	
	<i>a</i> o	6.2.2	Canonical Correlation Analysis	106 107
	6.3	Experiments		
		6.3.1	PET Imaging Data	107
		6.3.2	Non-Imaging Data	108
		6.3.3	Results	109
	6.4 Investigation of Stacking Performance		gation of Stacking Performance	111
		6.4.1	UCI Data Subsets Generation	111
		6.4.2	UCI Data Results	111
	6.5	Discus	sion and Conclusions	114
7 Conclusions and Outlook		and Outlook	116	
	7.1	Summ	ary	116
	7.2)k	117
Аp	pendi	ces		120
Ар	pendi	× A A	ppendix	121
	A.1	ADNI	Data	121
	A.2	Brain	Regions defined in AAL	123
	A.3	Deriva	tion of Expectation Maximization Algorithm	124
Lic	+ of E	igures		126
LIS	LOIF	igures		120
List of Tables				
List of Symbols 1				
Index 13				
Bibliography				

This page is intentionally left blank.

CHAPTER 1

Introduction

A part of patient data from medical application is stored in a structured manner. This data can come from different sources, such as brain scan images (for example, positron emission tomography), demographic information, neuropsychological tests, etc. Clinicians use diverse information to arrive at a diagnosis. However, clinicians may have difficulties in diagnosing some patients whose symptoms are ambiguous. Furthermore, clinicians can sometimes also disagree with each other. Therefore, gaining some insights into data not only help clinicians in diagnosis, but also assist them in making decisions. In general, we divide patient data into two categories, imaging data (e.g., PET scans) and non-imaging data (e.g., demographic information, neuropsychological tests). Therefore, we develop algorithms to discover hidden information in both data sources.

1.1 Mining Imaging Data

In this thesis, PET scans are used as the imaging data to investigate a patient's mental status in terms of brain metabolism activity. One application is in the area of dementia classification, i.e., in the attempt to devise a computer algorithm that classifies the class (NC, MCI or AD)¹ of a given PET scan based on a number of training samples. A different application may originate from a follow-up study.

¹ NC: normal control, MCI: mild cognitive impairment, AD: Alzheimer's disease. Definitions are given at Section 2.1.

2 1 Introduction

Given a PET scan from a patient with MCI, we try to predict whether this patient will develop AD within the next two years. The former is a study of the classification of images reflecting NC, MCI and AD, whereas the latter is a study of predicting a patient's progressive mental state based on the current MCI image.

1.1.1 Dementia Classification

To correctly classify a PET image into NC, MCI or AD is one of the most active research topics in computational medicine, because a correct diagnosis may help a patient in receiving a more targeted treatment. Approaches based on voxels² as features and based on statistical modeling represent the main methodologies. The latter differs from the former by taking voxel neighborhood effects into account. Our proposed approach is based on statistical modeling and clusters brain voxels into a certain number of clusters, where the number of clusters can be determined by a model selection criterion. Finally, the mean and standard deviation of a cluster are calculated as input features for another algorithm. The suggested Gaussian mixture model (GMM) with a model selection method shows better results than compared baseline methods, in particular for the MCI against AD case on two independent datasets.

1.1.2 Mild Cognitive Impairment Progression

The progression from MCI to full-blown AD is a great challenge which requires a different approach. Researchers have conducted studies on analyzing the brain scan difference between patients who progress to AD and those remain in the MCI state. Applying a simple t-test can yield some significant brain voxels that represent the interesting regions related to MCI progression. However, such a group comparison based method may have limited performance on new test data. To this end, we develop an algorithm that learns brain patterns of interest, which can be applied for MCI to AD progression prediction. The proposed survival analysis and infinite Gaussian mixture model (IGMM) method show some power in predicting MCI progression.

² Voxels are volume elements in three-dimensional space.

1.2 Mining Non-imaging Data

Non-imaging data can also be seen as structured data, which shows another perspective to patient data. Among the variables in these structured data, some particular variable combinations may indicate an interesting phenomenon. For example, male patients above the age of 65 tend to suffer from Alzheimer's disease. To find such a variable combination, subgroup discovery [Klö96, Wro97] can be employed, aiming at searching for variable combinations with respect to a given target class. Although subgroup discovery is a technique that has been around for more than twenty years, some open issues still remain, such as a huge search space or redundancy in the resulting rule set. In this work, two new methods [LAK14, LPDK15] are developed to efficiently discover subgroups, while keeping the redundancy at a low level.

1.3 Combining Imaging and Non-imaging Data

Today, it is common to combine different sources of information to reach a more reliable result. For example, multiple kernel learning, majority voting and stacking are candidate approaches, to name only a few. A stacked multi-view learning method [LHS+11] was devised, taking a decision based on a base and a meta model. The base level of the proposed approach collects decisions provided from the base level classifiers and the meta level classifier attempts to learn correlations between correct decisions and decision behaviors of base classifiers. The proposed method yields better results than using any individual source of information alone. In this way, it meaningfully combines information sources.

The thesis is structured in the following way:

Chapter 2 introduces dementia and Alzheimer's disease domain knowledge. We further outline the common diagnosis-relevant neuropsychological tests, automated diagnosis techniques and medical datasets used in the experiments.

Chapter 3 offers background knowledge on functional imaging techniques and related work in mining imaging data. Subgroup discovery is subsequently introduced in detail. Finally, we elaborate on multi-view learning techniques.

Chapter 4 is devoted to the mining of imaging data. A Gaussian mixture model with model selection is developed particularly for dementia classification. An MCI progression prediction study is then conducted using survival analysis and the infin-

1 Introduction

ite Gaussian mixture model.

Chapter 5 presents two new subgroup discovery algorithms. The optimization and topic modeling based algorithms show advantages compared to existing methods. Experiments are conducted on benchmark UCI datasets as well as a medical dataset.

Chapter 6 addresses the stacked multi-view learning that combines imaging and non-imaging data for a study. We also attempt to analyze the experimental results of stacked multi-view learning, revealing the factors associated with its performance.

Chapter 7 concludes the thesis and points out some feasible future studies.

This page is intentionally left blank.

CHAPTER 2

Background on Dementia and Alzheimer's Disease

2.1 General Introduction to Alzheimer's Disease

Alzheimer's disease is a progressive, degenerative and incurable disease of the brain and the main cause of dementia. According to studies from the World Health Organization (WHO), there are around six million patients suffering from Alzheimer's disease in North America alone. In Europe and China, the number even exceeds six million. The number of people suffering from dementia is expected to grow rapidly in the next decades. It is predicted that the number of patients with Alzheimer's disease will exceed 13 million in United States by the year 2050 [LZ14]. Worldwide, according to some study [ALZ08], the number is expected to exceed 100 million by 2050. This will also have a major negative impact on healthcare systems.

Alois Alzheimer first observed the Alzheimer's disease symptoms in 1901. The patient was Auguste Deter, who later died from Alzheimer's disease. Subsequent microscopic analysis of her brain regions revealed several phenomena. For example, noticeable atrophy of the cerebral lobes; loss of neuronal cells in some brain regions; the existence of strange fibrillary pathology inside the neuronal cells; presence of fibrous glial cells in the brain as well as in blood vessels. These effects continue to represent the pathological characteristics of AD today. Fig. 2.1 shows that the brain with Alzheimer's disease severely suffers from lost brain tissue, whereas the healthy brain remains intact.

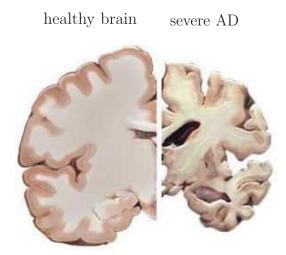


Figure 2.1: Healthy and Alzheimer's disease brain. Image courtesy of National Institute on Aging—National Institutes of Health.

Alzheimer's disease progresses in several phases, affecting memory, learning, judgement, emotions and even movement. In the first phase, the damage is not obvious and shows no visible symptoms. Later, some observable behaviors may appear, such as irritability, apathy, lack of interest, etc., but the behavior difference may vary from one to another. The brain regions and their respective functionalities are depicted in Fig. 2.2, showing several regions related to Alzheimer's disease. The parietal lobe is associated with body awareness, so people whose parietal lobe is in atrophy may suffer from losing control of the body. The temporal lobe is essential for long term memory, facial recognition and similar daily behaviors, thus it is also an important region in Alzheimer's disease. The cerebellum, on the other hand, controls balance, and it is widely regarded as a less affected region in Alzheimer's disease. Therefore, the cerebellum region is often used as a reference region in PET scan intensity normalization, because a reference region needs to be a healthy region that can be treated as a benchmark to compare with.

One observed fact is that the number of female patients with Alzheimer's disease is twice as high as the one of male patients. Noticeably, cardiovascular disease is known to be associated with Alzheimer's disease. Many scientists believe that it could explain the imbalance between male and female patients with Alzheimer. Because cardiovascular disease affects men in their forties or fifties, estrogen and related hormones in women help them get through this phase more easily. As a consequence,

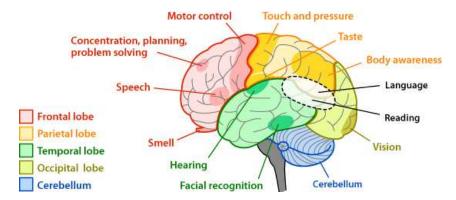


Figure 2.2: Brain regions and functionality. Image obtained from https://askabiologist.asu.edu/what-your-brain-doing.

men are at a higher risk of death in the decades before we can observe the presence of Alzheimer's disease. As for women, they may readily survive from cardiovascular disease at ages forties or fifties, but will likely be affected by Alzheimer's disease in their sixties or seventies [PG14]. However, the AD risk is higher in women versus men even after adjusting for higher life expectancy in women. Therefore, there must be other characteristics that protect men from AD, but we do not yet know what these are.

Age is considered the main risk factor of Alzheimer's disease. Senile plaques, neurofibrillary tangles and the massive loss of brain neurons are the three main biological indicators of Alzheimer's disease. Family history, obesity and head injury can also play a role in the development of Alzheimer's disease.

Apolipoprotein E4 allele (ApoE E4) is one important risk factor in Alzheimer's disease, since it plays an essential role in the cardiovascular system that transports and delivers blood cholesterol. In addition, smoking, alcohol abuse and depression may also cause Alzheimer's disease.

The educational level seems to play an important role in protecting people from getting Alzheimer's disease [Kat93]. Besides, some studies suggest that consuming red wine at a moderate level can provide some protection against Alzheimer's disease. Physical exercise may postpone the symptoms of Alzheimer's disease as well. Besides, intellectual exercise, such as chess or Sudoku, gives people the opportunity to improve memory or decision making, thus delaying the onset of the disease.

Concerning dementia and Alzheimer's disease, the terms can often be confusing

Alzheimer's disease Parkinson's disease
mixed cognitive impairment Frontotemporal dementia (FTD)
vascular dementia normal pressure hydrocephalus
mixed dementia Huntington's disease
dementia with Lewy bodies Wernicke-Korsakoff syndrome

Table 2.1: Various forms of dementia.

or misused, thus a new lexicon [DFJ⁺10] was defined.

Some definitions:

- Normal control (NC) refers to the person whose mental status is normal.
- Mild cognitive impairment (MCI) is the patient who suffers from subtle but measurable memory disorder. MCI is a term describing a syndrome, rather than a disease.
- Alzheimer's disease (AD) refers to the entire clinical phase of the disease and is not bounded by the syndrome of dementia [DFJ⁺10]. There are several terms describing different statuses of AD, for example, prodromal AD, AD dementia, typical AD, atypical AD, mixed AD.

Dementia is a general term that is described in Table 2.1, and Alzheimer's disease remains the main disease in dementia. The ICD-10, international classification of diseases, defines Alzheimer' disease as F00.0, F00.1 and F00.2. MCI is denoted as F06.7. These terms are used by clinicians in dementia diagnosis.

Categories	Subtests (0-15)
semantic memory	Boston naming test (0-15)
word finding	word list learning $(0-30)$
visual cognition	word list cognition $(0-20)$
orientation	word list recall $(0-10)$
concentration	construction praxis $(0-11)$
direct retentiveness	constructional praxis recall (0-11)
visuo-construction	verbal fluency $(0-\infty)$
delayed retentiveness	Mini-Mental State Examination (MMSE) (0-30)

Table 2.2: Categories and subtests in CERAD.

2.2 Neuropsychological Tests and Computer-aided Diagnosis

Neuropsychological tests provide clinicians with essential diagnostic information, reflecting various daily abilities like memory, orientation, etc. We review the commonly used Consortium to Establish a Registry for Alzheimer's Disease (CERAD) [MMR88], Clock Drawing Test (CDT), ADAS: Alzheimer's Disease Assessment Scale and Clinical Dementia Rating (CDR).

CERAD encompasses several daily abilities described by Table 2.2. MMSE is a very important subtest of CERAD, acting often as an indicator of mental health status. ADAS is another test including word recall, naming, commands, constructional praxis, to name only a few.

Besides question answering, patients can also be asked to draw a clock showing ten past eleven, with grade from one to six and one for a perfect clock, cf. Fig. 2.3.

As a contrast to other neuropsychological tests, CDR is much closer to a diagnosis. Thus, clinicians often use it as an important reference for diagnosis. It includes categories such as memory, orientation, judgement, community, activity, personal care and a global score. The possible scores are 0, 0.5, 1, 2, 3, with 0 as the best and 3 as severe.

In the year 1984, the criteria for clinical diagnosis of Alzheimer's disease were established by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Asso-

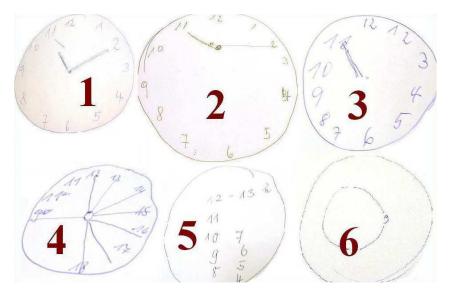


Figure 2.3: Example of clock drawing test. Image obtained from Master thesis [Sch07].

ciation (ADRDA). The current diagnosis of Alzheimer's disease relies heavily on biopsy or postmortem examination of the brain [LZ14]. In the meantime, the use of PET, computer tomography (CT), cerebrospinal fluid (CSF) and magnetic resonance imaging (MRI) provide a better understanding of brain volume loss and cognitive degradation. Despite technological progress, the *ante mortem* diagnosis of Alzheimer's disease is still based on clinical grounds, with biomarkers such as cerebrospinal fluid (CSF) proteins and neuroimaging procedures providing supporting information. In the following, we review some computer-aided diagnosis (CAD) approaches in diagnosing Alzheimer's disease.

Visual examination of images may be error-prone, thus CAD may be another tool to diagnose Alzheimer's disease [GLR⁺09, Met99]. One technique often used for CAD in this area, automatic image classification, remains a topic receiving intensive research. Overall, however, more studies have been conducted on MRI than on PET images. Regional brain atrophy, especially of the mediotemporal lobe, is a typical feature of AD, which can be reliably identified by MRI [dLML⁺07, FWFN⁺10, MFNR⁺09], and MRI is therefore a useful imaging biomarker. For example, a recent study [CGT⁺11] compared ten distinct MRI classification approaches using 509 subjects of the Alzheimer's Disease Neuroimaging

Initiative (ADNI), investigating the differentiation between different groups of individuals including normal controls (NC) against patients with AD. Another work [LZS12] presented an ensemble approach to combining a number of weak classifiers for classification. This local patch-based (a patch is understood as some small region) subspace ensemble method builds individual classifiers based on various subsets of local patches and then combines them for a better classification. A recent study [LZStADNI14] proposed a hierarchical tree to capture relationships among imaging features, aiming at identifying informative biomarkers for classifications of MCI and AD using MRI scans. Another study [KSC⁺08] investigated the use of support vector machines (SVM) for automatic classification based on MRI scans. It showed that SVM can be very helpful in diagnosing various forms of dementia, revealing that the linear kernel is adequate and generalizes well. A work [RCG⁺09] applied random forest classifier on SPECT images, showing very good classification accuracy. In recent years, multi-modality classification has been shown to be an attractive research area in Alzheimer's disease research. A pairwise similarity measure derived from random forests was proposed as a multi-modality classification framework [GAH⁺13]. In this study, FDG-PET, MR, CSF biomarker and categorical genetic data were employed for the classification. The results indicate that joint information is superior to any individual modality on its own. In another work, MRI data, PET data and CSF biomarkers were used to construct a kernel matrix, and a combined kernel was produced for the final classification [ZWZ⁺11]. This method allows combining heterogeneous data and permits different weights for various data modalities. The results show high AD classification accuracy, even in very early clinical stages (i.e., for mild cognitive impairment).

Combining imaging data and non-imaging data has also been studied. A work [SBS⁺] combined PET images and neuropsychological test for Alzheimer's disease classification. The experiment using 46 subjects revealed that neuropsychological tests can improve the classification accuracy compared to using the imaging data alone.

2.3 Data Used in Experiments

2.3.1 Medical Data

Two independent medical datasets are used in this work. One is an in-house data set provided by the Technische Universität München (TUM) School of Medicine. Patients' names were anonymized in the database. The study was approved by the ethics committee of the "Medical Faculty of Technische Universität München", which permitted the use of patient data. The data includes PET scans, CERAD, CDR, CDT and Apolipoprotein E, but some of the data records were missing for certain patients. The other dataset is from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is publicly available. ADNI offers more types of neuropsychological test data, such as ADAS, and even some other biomarkers, but there are often missing records. Some of the patient PET scans are adequately followed up at least two years, so we can perform a follow-up study to investigate Alzheimer's disease progression, cf. Section 4.2.

2.3.2 Benchmark UCI Data

UCI (University of California, Irvine) benchmark datasets are widely used in the machine learning and data mining communities [FA10]. The data is publicly available at https://archive.ics.uci.edu/ml/index.html. In the experiments, data records with missing values were removed.

This page is intentionally left blank.

CHAPTER 3

State-of-the-art in Data Mining

3.1 Mining Imaging Data

3.1.1 Introduction to Functional Imaging Techniques

PET is a non-invasive imaging technique, measuring blood flow, membrane transport and metabolism activity quantitatively. PET gives us detailed insights into molecular processes occurring in tissues. It is a three-dimensional medical imaging technology based on the detection of positrons. The detection of positrons is realized by a tracer that is injected into the body. There are several commonly used tracers, such as fluorodeoxyglucose, acetate, palmitate, hydroxyephedrine, ammonia, water and rubidium. Throughout this work, we use fluorodeoxyglucose type PET (FDG-PET) scans for our study. Today, radiology and nuclear images are often used in clinical practice. We briefly highlight the difference between these two types of images in Table 3.1.

Many studies have been conducted using MRI to investigate Alzheimer's disease, since MRI can show the atrophy of the entire brain of patients with Alzheimer's disease, especially in the hippocampus region. Different from the MRI, the FDG-PET reflects the decrease of metabolism in the posterior regions and in the posterior cingulum. The FDG-PET has been widely applied to assist the diagnosis of AD, because it has been shown that the brain scans can be very helpful in diagnosing Alzheimer's disease at an early stage [Drz09]. Besides, FDG-PET was also recommended as a diagnostic marker by recently proposed guidelines [DFJ⁺07]. A

Radiology	Nuclear Medicine
anatomical images	physiological images
structural images	functional images
macroscopic imaging	molecular imaging
ultrasound XR CT MRI	planar imaging SPECT PET

Table 3.1: Difference modalities in radiology and nuclear medicine imaging.

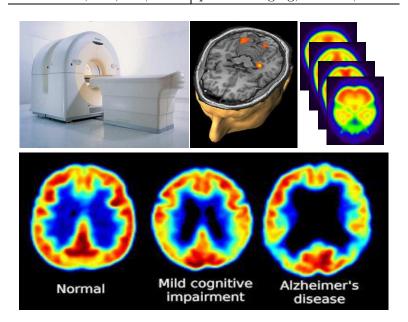


Figure 3.1: PET scanner, 3D brain, PET slices and NC, MCI, AD PET image. Bottom PET images courtesy of Suzanne Baker, PhD; William Jagust, MD; and Susan Landau, PhD.

significant functional neuroimaging application is the progression prediction of prodromal AD, i.e., mild cognitive impairment (MCI), to full-blown AD dementia. A growing number of studies show that FDG-PET can predict the clinical outcome in MCI with a relatively high sensitivity and specificity. A recent literature review underpins the gain in overall diagnostic accuracy by using FDG-PET in the evaluation of dementia, supporting its role as an effective complementary tool [BHAM12]. The superiority of FDG-PET to other potential predictors of clinical decline in MCI has been suggested by some [Lan10], but not all studies [Gom14]. Typically, a couple of AD-like FDG-PET patterns is found to be predictive of future AD dementia in patients with MCI [MTH⁺08], such as decreases of the cerebral glucose metabolism

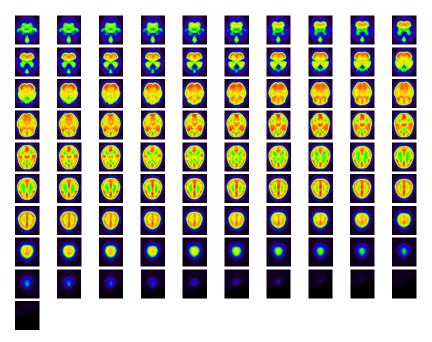


Figure 3.2: 91 PET scan slices in transaxial view. The colormap is defined as rainbow in XMedCon. Red color represents intense activity and black is no activity at all. The black region represents no brain region, i.e., a non-brain part.

in the posterior temporoperietal regions and in the posterior cingulate cortex, as well as in the prefrontal cortex.

Visualization offers a straightforward way to inspect the brain scans. Fig. 3.2 depicts 91 slices in transaxial view. In this work, we use XMedCon, MRIcron to visualize and pre-process images. The 3D PET scan can be viewed in three different perspectives, namely, the coronal view, the sagittal view and the transaxial view, cf. Fig. 3.3. The 3D coordinate system consists the left-handed system (left cerebral heimsphere is at left) and the right-handed system (left cerebral heimsphere is at right). The right-handed system specifies -x coordinates are on the left and all +x coordinates are on the right, -y coordinates are on the posterior and all +x coordinates are on the anterior, -z coordinates are on the superior and all +z coordinates are on the inferior. By convention, a left-handed system is used in radiology, and a right-handed system is used in neurology. The PET images need to be pre-processed by SPM with the following steps:

• Spatial normalization: This ensures individual images are mapped onto a

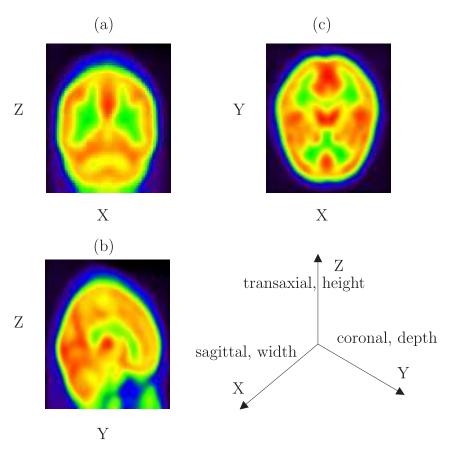


Figure 3.3: Three views of PET scan and coordinate system. (a): coronal view (b): sagittal view (c): transaxial view. P: primary sensorimotor cortex normalization; G: grand mean normalization. x, y and z are the width (91), depth (109) and height (91) respectively.

standard brain template, so that the brain region analysis is based on a common basis. The dimension size (x,y,z) of the resulting image can be adjusted by the specified bounding box in SPM. We used the bounding box $[-90\ -126\ -72;\ 90\ 90\ 108]$ to obtain an image size of $91\times 109\times 91$, which is the same as the automated anatomical labeling (AAL, cf. Section A.2) [TMLP+02] brain template.

- Smoothing: This is commonly used in image processing for improving the signal to noise ratio by applying a fixed smoothing window to the image.
- Intensity normalization: This is an essential step in analyzing PET scans. Be-

cause patients can be injected with different amounts of tracer, the absolute brain metabolism activity may vary as a consequence. Thus, it is necessary to obtain the relative brain metabolism activity, which can be done by dividing each voxel to a value calculated from a chosen reference region. There are a couple of common choices in selecting the reference region. First, the mean intensity of the whole brain is known as the grand mean normalization. Second, the cerebellum region can be used, because it is not much affected by Alzheimer's disease. The third method may be a pSMC-based (primary sensorimotor cortex) approach, which was shown to perform well in a study [YLB⁺08].

3.1.2 Related Work in Mining Voxel Data

Data mining methods have been used in many applications in dementia research. A survey [MFS⁺00] reviewed the data mining methods used for brain imaging data and other clinical data. In another study [SF05], SPECT images were used to classify patients with AD based on a linear programming formulation. Clustering and subgroup discovery were studied to incorporate imaging and non-imaging data, where the PET images and clinical variables were combined to correlate disease patterns of the brain with neuropsychological tests [SHM⁺10]. A related work [GGK⁺06] applied stacking to the early diagnosis of AD against NC using event related potentials (ERPs) and showed that it outperformed majority voting. In the following, we specifically address two main methods in mining voxel data.

3.1.2.1 Voxel-based Morphometry (VBM)

Voxel-based morphometry (VBM) [AF00] has been a popular approach to analyzing functional imaging data, aiming at finding group differences in brain regions. Statistical parametric mapping (SPM) provides VBM as a fundamental analysis technique, which performs statistical tests, such as the t-test, across different groups of subjects to discover discriminative voxels. Specifically, as for MRI scans, images are first spatially normalized using a standard brain template, such that further comparison and analysis can be performed on the same basis. Images are subsequently segmented into gray matter, white matter and CSF, followed by image intensity normalization. Then, image voxels are smoothed by taking the surrounding voxels

into consideration. Afterwards, statistical analysis plays an important role in inferring the statistical difference existing in various groups. In SPM, a wide range of analysis methods are provided, such as the t-test, ANOVA (analysis of variance), etc. Because false positives may appear due to statistical tests performed on a large number of voxels, it is usually recommended to apply multiple comparison correction. A large family of such methods is available, for example, family-wise error (FWE), false discovery rate (FDR) and Bonferroni correction. In terms of PET scans, the procedure is similar, except that PET scans need not to be segmented into gray matter, white matter and CSF, since PET scan is essentially a reflection of gray matter metabolism activity.

The VBM approach is straightforward. Many studies have been conducted using the voxel-based analysis methods. For example, voxels were directly employed using an SVM for classification in a study [KSC⁺08]. However, such an approach ignores statistical association among voxels. A couple of weaknesses were mentioned in a related work [Tha08]. For example, the assumption of one-to-one mapping among human brains may not be meaningful. Indeed, a VBM approach is quite sensitive to the individual differences. Further, the statistics applied is criticized because the identified regions may suffer from a misregistration from the previous spatial normalization step.

3.1.2.2 Statistical Approaches to Voxel Mining

The voxel-based approach suffers from considering voxels independently, which takes no correlations between neighbouring voxels into account. In contrast to such a univariate approach, two multivariate methods were proposed to learn the patterns in data from a higher perspective. In the first study, principal component analysis (PCA) was used to extract features, which were then fed into a classifier [LRG⁺09]. Another work [NSM⁺08] used PCA to analyze FDG-PET in amnestic MCI and illustrated some interesting findings using the principal components. A combined use [LRG⁺] of kernel principal component analysis (KPCA) and linear discriminant analysis (LDA) was applied on SPECT images to extract features. This method outperformed a voxel-based feature selection approach by an accuracy gain of 12%. It is worth mentioning that the PCA based method used in the above cases essentially transforms the original features into another feature space, which is different from the Region of Interest (ROI) based method. ROI based methods derive fea-

tures directly from the PET images [GSR⁺11, SGR⁺12]. A t-test feature selection with feature correlation weighting was employed to form ROIs that were used for defining normalized mean squared error (NMSE) features [RCG⁺09]. A Gaussian mixture model [GSR⁺11] models the difference between the controls and patients with Alzheimer's disease, where the number of Gaussians (K) was fixed to 64, which can be a drawback since 64 may not be the optimal value.

3.1.3 Model Selection in Neuroscience

We also review some related work on model selection based on PET images in areas other than dementia research. In a PET volume classification study [SAAZ12], Bayesian information criterion (BIC) was applied to select the optimal number of classes for each PET scan. In this work, the BIC values gradually reached a steady state, such that the optimal number of classes was easily chosen. Another work [SCA⁺14] employed the Akaike information criterion (AIC) to assess the different predictive models, investigating the overall survival in a phase II clinical trial of a targeted therapy. The authors reported that the highest prognostic value appeared with the lowest AIC value, which suggested that AIC can be a guideline in choosing a desired model. In a proton therapy research study [EMC⁺13], the AIC was used to determine the most appropriate model for the FDG uptake dose response for each patient. A compound-B based PET kinetic modeling study [LKTR12] used AIC to ensure a good kinetic parameter setting. In another study [PKL⁺05], however, both AIC and BIC did not perform reliably for realistic 3D dynamic PET images. The authors assumed that the reason for this may be that AIC and BIC are model dependent, so the specified probability distribution function was not suitable for realistic 3D dynamic PET images. Although AIC and BIC are frequently applied model selection techniques in neuroscience, their application to PET scans for the purpose of AD diagnosis is rarely studied. Hence, this work studies their usefulness in combination with the proposed GMM approach in dementia research.

3.2 Subgroup Discovery

Subgroup discovery (SD), one field in pattern mining, aims to find interesting subgroups (conjunctions of variables) with respect to a target class. Subgroup discovery is closely related [NLW09] to other techniques, such as emerging pattern mining

[DL99], contrast set mining [BP01, NLW09], class-correlated pattern mining [ZR04] and rule learning [CS99, FP08, RK08].

Emerging patterns (EPs) [DL99] are defined as itemsets whose support increases significantly from one dataset to another. Contrast set mining [BP01] studies the differences existing in different contrasting groups by looking at conjunctions of attributes³ and values that show different supports. A recent survey [NLW09] discusses commonalities and differences. CorClass (correlated classification) [ZR04] is an approach for class-correlated pattern mining, which maximizes convex correlation measures such as information gain or χ^2 , as also proposed previously by Sese and Morishita [MS00]. All these approaches differ from SD (a) in the representation (propositional or relational rules vs. patterns from some pattern language like itemsets, trees, graphs) and (b) in their purpose (discovering interesting patterns, either for direct inspection by the user or for combining them into predictive models). However, search and pruning techniques developed there can in principle also be used for subgroup discovery, if the quality function fulfills properties like convexity.

Rule learning has been investigated in machine learning and data mining for more than forty years. SLIPPER [CS99] repeatedly boosts a simple and greedy rulebuilder to generate an ensemble of rules that is highly predictive. Much like SLIP-PER, examples are re-weighted after each iteration in our approach. Margin-based first-order rule learning [RK08] combines rules from an existing repository of rules by optimizing the Margin Minus Variance (MMV) criterion. Error bounds are used for capacity control. Here, numerical optimization is used to integrate rules into a weighted rule set. A rule ensemble framework [FP08] constructs classification models by linear combinations of simple rules yielded from the data. The principal advantage resides in its simple form for easy interpretation. Whereas rule learning and subgroup discovery share the same representation (rules), they differ in their goal to build a global or local model, respectively. In contrast to other approaches that use numerical optimization for rule learning, the approach presented here does so in a pre-processing step, to select subsets of relevant features that can be handed over to a traditional search through a lattice. The following sections discuss subgroup discovery, related algorithms and existing challenges.

³ Attribute and feature are used interchangeably in the thesis.

3.2.1 Introduction to Association Rule Mining

We review association rule mining as some background information, due to the close relation between subgroup discovery and association rule mining. Association rule mining [AIS93] and pattern mining aim at finding interesting regularities (patterns, rules) in a database. A typical application is the study of transaction data records in a supermarket, which may reveal a hidden rule that people often buy beer and chips together. This is actionable knowledge, since supermarkets can put them close to each other for the convenience of customers' purchase. To quantitatively measure the usefulness (interestingness) of a rule, researchers have introduced some measures, such as support and confidence:

$$support(A \Rightarrow B) = P(A \cup B). \tag{3.1}$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)}.$$
 (3.2)

Support can be regarded as a global measure of a rule, whereas confidence is a local measure of a rule that is irrespective of the total number of observations. For example, 1000 customers have visited a shop, 400 of them bought computers and 50 of them bought software as well. Therefore support($computer \Rightarrow software$) = $\frac{50}{1000} = 5\%$ and confidence(software | computer) = $\frac{50}{\frac{1000}{1000}} = \frac{50}{400} = 12.5\%$, which shows that the confidence is independent from the total number of customers.

In the context of association rule mining, coverage is also called support (frequency), accuracy is also called confidence. However, support in fact means the accuracy in the subgroup discovery literature. We follow the convention in their contexts respectively. Support and confidence can be misleading, thus they are insufficient at filtering out uninteresting association rules. To cope with it, correlation analysis may be helpful. To this end, lift is used as a measure of correlation between A and B:

$$\operatorname{lift}(A,B) = \frac{\operatorname{confidence}(A \Rightarrow B)}{P(B)} = \frac{P(A \cup B)}{P(A)P(B)}.$$
 (3.3)

If lift is less than one, then A and B are negatively correlated. A and B are positively correlated if lift is greater than one. They are independent if lift equals one. A second correlation measure is the χ^2 (independence) measure. However,

the χ^2 measure is not anti-monotonic with respect to set-inclusion. Besides, most statisticians warn against using the test when any of the expected values is less than five. Instead, Fisher's exact test is suggested. In addition to lift and χ^2 , all-confidence and cosine analyses are recommended to use when the result shows that weak correlation is present, then other analysis may be performed to obtain a more complete picture [HK05]. In the following, we highlight some important concepts:

- An itemset is *closed* if none of its proper super-itemset has the same support as the itemset.
- An itemset is *maximal* frequent if none of its super-itemset is frequent. If an itemset is maximal it is also closed.
- An itemset is *derivable* if its support can be exactly inferred from the support of its sub-itemsets based on inclusion-exclusion principle. Otherwise it is non-derivable. A family of all non-derivable itemsets is downward closed.
- The Apriori property tells that all nonempty subsets of a frequent itemset must also be frequent. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1)-itemsets. This property belongs to a special category of properties called anti-monotone in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well. It is called anti-monotone because the property is monotonic in the context of failing a test. Anti-monotone: given $X \subseteq Y$, if C(X) is not true then C(Y) is not true, i.e., $\neg C(X) \to \neg C(Y)$. Monotone: given $X \subset Y$, if C(X) is true then C(Y) is true, i.e., $C(X) \to C(Y)$. Anti-monotone is useful for bottom-up searching, while monotone is useful for top-down. The apriori property plays an important role in subgroup discovery, since it offers a methodology to find subgroup rules.

In subgroup discovery, rules are found with respect to a target variable (class, label), whereas there is no target variable in association rule mining. For example, a rule can be an association rule if it is written as "buy beer and chips together". A slight modification leads it to a subgroup discovery rule if it becomes "IF buy beer and chips together, THEN the customer is male". Male is a target variable, beer and chips are conjunctions of conditions pointing to the target. The subsequent section introduces subgroup discovery in more details.

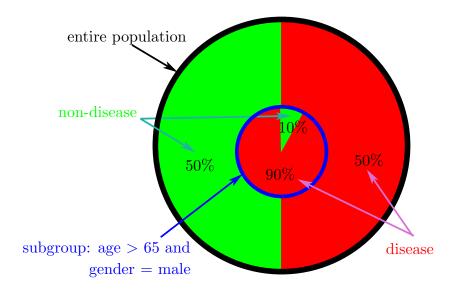


Figure 3.4: Pie chart illustration of subgroup discovery.

3.2.2 Introduction to Subgroup Discovery

Subgroup discovery is a task at the intersection of predictive and descriptive induction, aiming at identifying subgroups that have the most unusual statistical (distributional) characteristics with respect to a property of interest. The task of subgroup discovery (SD) is to find population subgroups described by conjunctions of attribute-value conditions that are statistically most interesting (e.g., large, but at the same time distributionally unusual) with respect to a property of interest [Klö96, Wro97]. It is considered a task at the intersection of predictive and descriptive induction. An example of a subgroup rule could be: "IF age > 65 AND gender = male, THEN disease", where "age" is an attribute (feature), "> 65" is a condition of this attribute, and "disease" is the target class. Subgroup rules have conjunctions of conditions on the left-hand side and a user-specified target class on the right-hand side.

In Fig. 3.4, 50% of the entire population has a disease, but the percentage dramatically increases to 90% in a subgroup "age > 65 and gender = male". Thus, we may be interested in such a subgroup representing obvious contrast to the whole population. Quantitatively, the interestingness of a subgroup can be measured by a quality function.

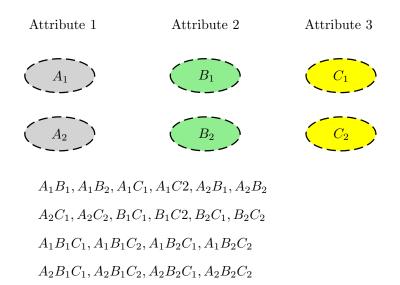


Figure 3.5: Example of combinations of attribute-value conditions in finding subgroup discovery rules. Every attribute has two distinct values. In practice, the attributes may be discretized if the data is continuous.

The process of subgroup discovery needs exhaustive search of all possible combinations of attribute-value conditions. As demonstrated in Fig. 3.5, three attributes with six distinct values produce 20 combinations to be inspected. Specifically, any possible attribute-value combinations (two attributes, three attributes) needs to be investigated. The number grows dramatically fast as more distinct values are present. Section 3.2.5 summarizes this issue in more detail.

3.2.2.1 Quality Function of Subgroup Discovery

SD is usually evaluated by a quality function providing a trade-off between rule generality and distributional unusualness. The most common form is:

$$\rho = g^a(p - p_0),\tag{3.4}$$

where $0 \le a \le 1$, where g is the generality (coverage) of the subgroup, p is the rule accuracy (support), i.e., the fraction of rows of the target class in the subgroup, and p_0 is the default rule accuracy, i.e., the fraction of rows of the target class in the database. We use Fig. 3.4 as a numeric example, if we let a = 1 and

g=0.15 (suppose the area of the inner pie accounts for 15% of the entire pie), then the quality of subgroup "age > 65 and gender = male" can be calculated as $\rho=0.15\cdot(0.9-0.5)=0.06$. If a quality threshold is set to 0.01, then 0.06 > 0.01 qualifies such a subgroup.

When a=1, Eq. 3.4, known as Piatetsky-Shapiro [Klö96] quality function, is equivalent to the weighted relative accuracy (WRAcc [LKF⁺04]) that can be expressed as:

$$WRAcc(Class \leftarrow Cond) = p(Cond)(p(Class \mid Cond) - p(Class)), \qquad (3.5)$$

where "Cond" represents conjunctions of attribute-values. To avoid some instances (samples) being covered by rules over and over again, a weighted covering algorithm was used to decrease the weights of the used instances [LKF⁺04].

$$\rho = \frac{n'(Cond)}{N'} \left(\frac{n'(class, cond)}{n'(cond)} - \frac{n'(class)}{N'} \right), \tag{3.6}$$

where N' is the sum of the weights of all instances, n'(Cond) is the sum of the weights of all covered instances. A tighter optimistic estimate (TOE) was proposed as $gp(1-p_0)$ [GRW08], as opposed to $g(1-p_0)$ [Wro97].

3.2.3 Subgroup Discovery Algorithms

We review some important works. Besides the following works, other related studies are discussed in a comprehensive overview [HCGJ11].

- 1. The system EXPLORA [Klö96] was the first approach for subgroup discovery, introducing subgroup discovery in a single-relational setting. Both exhaustive and heuristic search strategies are possible to be applied.
- 2. MIDOS [Wro97] extended EXPLORA to the multi-relational setting, using the concepts of the optimistic estimate and minimal support pruning, an optimal refinement operator as well as sampling to ensure efficiency. Two pruning strategies were proposed: (a) minimal support pruning: Because the descendant of a hypothesis cannot cover more instances than the hypothesis itself, the subtree can be entirely pruned once this hypothesis fails to meet the criterion.

- (b) optimistic estimate pruning: It can be applied to shrink the search space by removing branches that cannot pass a quality threshold.
- 3. SubgroupMiner [KM02] was the first algorithm to consider continuous target variables via discretization. It is an extension to EXPLORA and MIDOS, employing several quality measures to gauge the interestingness of discovered subgroups.
- 4. Scholz [Sch05] proposed a knowledge-based sampling method, offering a generic way of incorporating prior knowledge in the sense of sampling. They claim that knowledge-based sampling allows to construct small sets of rules with high diversity, revealing different aspects of a dataset.
- 5. CN2-SD [LKF⁺04] adapts CN2 [CN89], a standard classification rule learner, to subgroup discovery, with two distinct features: (a) the weighted relative accuracy measure trading off generality for accuracy, and (b) using weights for examples in the covering process.
- 6. SD-Map [MF06] proposed a fast but exhaustive subgroup discovery algorithm. It incorporates the idea of FP-growth, and is also able to handle missing values. Several quality functions like Piatetsky-Shapiro [Klö96], unusualness or the binomial test can be applied in this approach.
- 7. A tight optimistic estimate [GRW08] was proposed to limit the search space size. The tight optimistic estimate (TOE) is a tight upper bound of the quality of the promising subgroups. It is an improvement of the optimistic estimate [Wro97] that was shown to be not tight.
- 8. A ranking support vector machine approach was suggested to rank subgroups with respect to a user's concept of interestingness, being one of the few approaches that incorporate numerical optimization into subgroup discovery [RÖ9].
- 9. Difference-based estimates for a generalization-aware (DBEGA) [LBP13] method was presented to take generalization into account using some new optimistic estimates bounds. The method evaluates subgroups by considering the difference in pattern generalization. It is shown to outperform compared approaches by up to an order of magnitude in terms of runtime.

10. More recently, an interactive subgroup discovery approach was proposed to allow user feedback during search, obtaining more interesting and diverse rules [DL13].

3.2.4 Evaluation Measures

Subgroup discovery can be evaluated by different measures [LKF⁺04]. Some commonly used ones are as follows:

• Cover redundancy (CR) [LPDK15, vLK11] measures the cover count of each sample covered by the rule set, and the deviation from the mean cover count is used to judge the level of redundancy. If the rule set covers some samples unevenly and ignores the others, then this rule set focuses too much on one part of the data. Hence, it probably has some degree of redundancy. Therefore, a lower CR suggests the subgroup rule set covers the data fairly well and is less redundant. Denote a dataset as \mathcal{D} and a set of subgroups \mathcal{S} . The cover count (CC) of a sample m is simply how many times this sample is covered by the rule set \mathcal{S} , i.e., $CC(m,\mathcal{S}) = \sum_{s \in \mathcal{S}} \mathcal{D}_s(m)$. The expected count $\overline{CC} = \frac{1}{|\mathcal{D}|} \sum_{m \in \mathcal{D}} CC(m,\mathcal{S})$. The CR is then computed as:

$$CR^{\mathcal{D}}(\mathcal{S}) = \frac{1}{|\mathcal{D}|} \sum_{m \in \mathcal{D}} \frac{|CC(m, \mathcal{S}) - \overline{CC}|}{\overline{CC}}.$$
 (3.7)

The CR is supposed to compare different subgroup sets of (roughly) the *same* size for the same dataset [vLK11].

• Jaccard index (JI) is employed as a measure of the diversity of a rule set. Given rules r_1 and r_2 from a rule set \mathcal{R} , the JI is calculated as:

$$JI(r_1, r_2) = \frac{|r_1 \cap r_2|}{|r_1 \cup r_2|}. (3.8)$$

The rules have common elements only when they have matched feature values. As JI (the lower, the more diverse) is computed in a pair-wise manner, we compute it for every two rules in the rule set \mathcal{R} . Then the mean JI is $\frac{\sum_{i=1}^{n} JI_i}{n}$, where $n = \binom{|\mathcal{R}|}{2}$ is the number of comparisons.

• Accuracy reflects the predictive power of the resulting rule set.

- *Number of rules* is related to the amount of time a human may need to examine and interpret the rules.
- Number of attributes per rule represents the number of attributes covered by a rule in a rule set. A rule with a single attribute is the simplest rule, with several attributes being more complex.

3.2.5 Challenges in Subgroup Discovery

The exponential search space and redundancy constitute two major challenges in subgroup discovery.

The search space in subgroup discovery grows exponentially, therefore exploring the full search space can be computationally very expensive. To cope with it, beam search [KM02] is often used as a greedy search approach to limiting the search possibilities. In addition, optimistic estimate [Wro97, GRW08] is another method to exclude unpromising search branches. Some other works are proposed in this regard as well. A closure system [BG09] was suggested to search for extensions of quality functions rather than their individual descriptions. The authors found that the search space and outputs were efficiently reduced when equivalence classes are considered. A random sampling method to obtain maximal itemsets was proposed, which can make use of any monotonically decreasing measure as an interestingness criterion [MG13].

Redundancy is an essential issue in pattern mining, which has been actively studied in recent years. The redundancy in solution sets remains one of the big open problems in subgroup discovery. The redundancy in resulting rule sets is partially caused by exploring huge search spaces. In the following, we review some related works concerning redundancy. Compared to the variety of different SD algorithms, research on redundancy in subgroups appears to be quite limited. Due to the close relationship between SD and itemset mining, it is worth mentioning work on avoiding redundancy with itemsets. A study [XCYH06] examines evaluation functions for measuring the combined significance of a pattern set, and proposes the use of MMS (maximal marginal significance), aiming for significant and non-redundant top-k patterns. In beam search, the top-k subgroups may contain variations of the same theme, whereas other interesting patterns may be left out. Besides, top-k mining has the drawback of considering each subgroup individually. To address this

issue, an approach that selects diverse subgroups instead of the top-k in conventional beam search was proposed, attempting to find non-redundant subgroups by a modified beam search [vLK11]. A memory-efficient, relevance test based top-k subgroup mining algorithm was the first to find the top-k relevant subgroups by a traversal of the closed-on-the-positives [GP11]. Another work [BZ09] on constrained pattern mining suggests two general heuristic algorithms, "Bouncer" and "Picker", to select a small subset of patterns. The reduced pattern set indeed improves the classification results. The MDL (minimum description length) principle has witnessed a renaissance recently. The KRIMP algorithm [VLS11] uses an itemset coding scheme to compress the data effectively. Related work employs a probabilistic maximum entropy model to iteratively find itemsets [MVT12]. The authors suggested the use of MDL to identify itemsets that summarize the data well. Most recently, diverse subgroup set discovery (DSSD) [LK12] attempted to obtain diverse rule sets by integrating a beam search into each level-wise search.

In Chapter 5, we present two approaches for subgroup discovery. One avoids redundancy neither during search nor by post-processing, rather it employs a quadratic programming approach as a pre-processing step. The other is a modified topic modeling based approach to finding subgroup rules.

3.3 Multi-view Learning

Today, it is common to have various information sources, which can be naturally studied by multi-view learning. Multi-view learning is based on the assumption that the views (various types of data) are both compatible and uncorrelated. One real-world multi-view application may be that a patient with dementia can be diagnosed by cognitive tests, brain scan or other biomarkers. In the following, we point out the related studies conducted for multi-view learning, which is then followed by a stacking method [Wol92] that can be used to learn the multi-view data.

3.3.1 Multi-view Learning Related Work

Co-training [BM98], a semi-supervised approach, is an early work on multi-view learning. For example, a web page has two views, i.e., the text appearing on the page itself and the anchor text attached to the hyper-links pointing to this page. Using these two information sources shows improved web page classification results.

A related work [CM09] developed a co-training algorithm to use data from various sources for prediction. The data sources constitute several assay screens describing individual molecules in a drug discovery application, reflecting both biological and chemical views.

Multiple kernel learning (MKL) is another framework to deal with multi-view learning, corresponding naturally to different views and combining kernels either linearly or non-linearly. Due to its closeness to kernel methods, one approach [BLJ04] formalizes MKL as a second order cone program problem, and a sequential minimal optimization (SMO) based SVM algorithm was proposed to obtain the optimal solution. Besides these theoretical works, a study [FRD11] employs MKL to integrate imaging and non-imaging features, which together predict the cognitive decline of the healthy adults. The results suggest that the joint information outperforms either individual data source. In computer vision, object detection [VGVZ09] benefits from MKL by capturing different feature sets, such as distribution of edges, dense visual words, spare visual words and feature descriptors. Multi-view clustering [BS04] was proposed to cluster text data, as contrast to other multi-view works devoted to classification. They found that the suggested multi-view k-means and EM algorithms performed better than their single-view counterparts. They also reported negative results, and applied entropy analysis to discover the underlying causes. A related work [CUD08] specifically focused on the issue of view disagreement in multi-view learning. A conditional entropy criterion was presented to detect view disagreement. Subsequently, detected samples causing view disagreement were filtered so that standard multi-view learning methods can be applied as usual.

Aside from above works, some other works are devoted to introducing other techniques in multi-view learning. One [DFU11] proposed a low rank multi-view learning (LR-MVL) method, which is a spectral method dealing with a low dimensional space. It has the advantages of being both computationally efficient as well as avoiding becoming stuck in local optima, thus is able to be trained using large amounts of data. Another study showed how to construct embedding projections from data, such that the Euclidean distance yields a meaningful similarity measure for both within-view and between-view. The experiments showed that the proposed nearest neighbor retrieval is feasible, exceeding the performance over principal component analysis and canonical correlation analysis as baselines. The use of both labeled and unlabeled data addresses transductive learning, which can be integrated into

multi-view learning [CmJ09]. The work introduced a linear smoother to represent each view, thus the multi-view transductive learning can be studied under a comprehensive generalized fixed point additive modeling framework. In this framework, the issue of view selection is handled by using a generalized AIC. The use of back fitting and local-scoring type algorithms enables an efficient implementation.

3.3.2 Stacked Multi-view Learning

In data mining, stacking (stacked generalization) [Wol92] has been widely used, but its underlying mechanisms are still not understood completely. An influential work [TW99] pointed out two interesting facts: First, multi-response linear regression (MLR) was a suitable meta level learner compared to decision trees, Naïve Bayes and K-nearest neighbors (KNN). Second, class probabilities (soft) should be used instead of class predictions (hard). The advantage of using class probabilities instead of class predictions was confirmed in further studies [SPSH05]. An interesting study argued that meta level correlation is crucial in stacking [FCS96]. Other ensemble methods are bagging and dagging [TW97]. Bagging employs joint sample bootstrapping, as opposed to dagging which uses disjoint samples. The extended new methods were named as bag-stacking and dag-stacking, indicating comparable results to bagging and dagging because no clear evidence suggested one being superior to the other. Additionally, it was demonstrated that stacking was better than selecting the best classifier from the ensemble by cross validation [Dz04]. More recently, stacked graphical learning was proposed for collective classification, showing that it was not only accurate but efficient as well [Kou07]. Voting and stacking were together studied in information extraction systems [SPSH05]. Both voting and stacking can benefit from using probabilistic estimates at the base-level. Stacking, compared to voting, was proven to perform better or at least comparable over all applications. Stacking was also shown to improve the prediction of membrane protein types [WYC06]. In this study, the SVM and instance-based learning were chosen as base level classifiers, and C4.5 was applied as a meta level classifier. As a result, the approach witnessed around a 20% overall accuracy gain.

Fig. 3.6 illustrates the framework of stacking. An information source (view or feature subset) contains test and training data, which are used to construct the meta level features. A base classifier is applied to the training data to yield the class

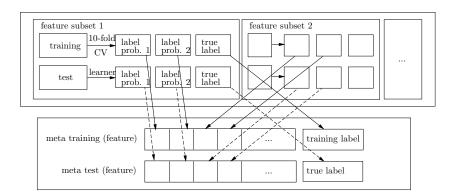


Figure 3.6: Demonstration of stacking. CV: cross-validation. feature subset: view, i.e., information source. prob.: probability.

probability estimates for both classes (assume two classes) using a 10-fold cross-validation. The yielded class probability estimates serve as meta level features for building the meta level training model. The meta level test feature is similarly constructed by applying the same base classifier to the test data. However, the true labels (if available) of the test data are reserved only for validating the final result. In such a manner, the meta level classifier attempts to learn the relationship between true decisions and the decisions made by various base classifiers.

This page is intentionally left blank.

CHAPTER 4

Mining Discriminative Imaging Features

4.1 Gaussian Mixture Model and Model Selection: A Classification Case Study

4.1.1 Motivation

Roughly speaking, the study of Alzheimer's disease has focused mainly on the pathological causes and the discrimination from other types of dementia or cognitive impairment. This study is dedicated to distinguishing mild cognitive impairment (MCI) from clinically diagnosable AD, MCI is a transitional phase to AD, which, however, does not necessarily end up in it. Hence, distinguishing MCI from AD is of great medical interest. Besides, classification cases as NC against AD and NC against MCI are also investigated. For this purpose and, generally, diagnosing different forms of dementia, imaging techniques like MRI and PET are used routinely.

Despite the compelling evidence in favor of FDG-PET as a prognostic marker, most experts concur that there is an immediate need for further efforts to improve implementation of neuroimaging in current diagnostic paradigms, including the improvement of the image analysis methods. Previous studies suggest that improved analytical methods such as principal component analysis [HFP+08], linear programming discriminant analysis [YLTS14], support vector machines (SVM), Gaussian process classification [YMC+13] or tree structured sparse learning [LZStADNI14] may improve the overall diagnostic and prognostic performance. We propose an

algorithm that is able to choose the optimal K by a model selection technique in Gaussian mixture model (GMM). Therefore, the proposed novel CAD approach can discriminate between NC, MCI and full-blown AD, which is often of practical interest.

4.1.2 Materials and Methods

4.1.2.1 Data Acquisition

Experiments were performed on two independent datasets. One is from the publicly available ADNI database (http://adni.loni.usc.edu/) and the other is an in-house dataset of patients and controls recruited at the memory outpatient unit of the Department of Psychiatry at Technische Universität München (TUM). The above datasets are in the following referred to as ADNI and TUM, respectively. ADNI has a large pool of PET (co-registered, averaged) images, which have been acquired on various scanners using different imaging parameters. To eliminate the bias of these factors, we selected images that have been obtained using the same scanner as well as the same parameters, such as the number of slices. The patient information and the PET scanner type are summarized in Table 4.1. Further details about the ADNI recruitment procedures are provided in the Appendix A.1.

Prior to their use for image analysis, PET images had to undergo two preprocessing steps: spatial normalization and smoothing (kernel size [8 8 8] mm), which were achieved by SPM5. The spatial normalization ensures that the processed image is of the size $91 \times 109 \times 91$, which is in accordance with AAL [TMLP⁺02]. The final step is the intensity normalization that was done by dividing each voxel by the mean intensity value averaged over all brain voxels (grand mean normalization, the non-brain voxels surrounding the brain were excluded). The second intensity normalization method is called pSMC (primary sensorimotor cortex) and was reported to be advantageous in a study [YLB⁺08]. Anatomically, the "Precentral_L, Precentral_R, Postcentral_L and Postcentral_R" regions in the AAL brain template can be used as the primary sensorimotor cortex.

4.1.2.2 Gaussian Mixture Model and Model Selection (GMM+MS)

GMM [Bis06] is a parametric density estimation approach that assumes the data is generated by more than one Gaussian distribution. It can cluster a point by assigning

Table 4.1: Participant characteristics and scanner type. NC: normal control, MCI: mild cognitive impairment, AD: Alzheimer's disease, TUM: Technische Universität München dataset, ADNI: Alzheimer's Disease Neuroimaging Initiative dataset, MMSE: Mini-Mental-State Examination.

		Subjects male:female	Age mean±std	MMSE mean±std	Scanner Type
ADNI	NC	30 (21:9)	74±5	28.6 ± 1.35	Siemens/CTI
	MCI	29(23:6)	74 ± 6	$27.4 {\pm} 1.68$	Siemens/CTI
	AD	25 (15:10)	72 ± 6	$23.2 {\pm} 2.23$	Siemens/CTI
TUM	NC	16 (7:9)	66 ± 6	$29.3 {\pm} 0.70$	Siemens Ecat HR Plus
	MCI	30 (16:14)	69 ± 7	$26.3 {\pm} 2.41$	Siemens Ecat HR Plus
	AD	30 (18:12)	69 ± 8	21.5 ± 5.23	Siemens Ecat HR Plus

the cluster identifier to the Gaussian that contributes the largest probability. Given a vector (data point) x, a GMM is defined as:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x; \mu_k, \Sigma_k), \tag{4.1}$$

where μ_k , Σ_k and π_k are the mean, covariance and mixing proportion respectively. In addition, $\Sigma_{k=1}^K \pi_k = 1$, $\pi_k \geq 0$ and $\theta = \{\mu_k, \Sigma_k, \pi_k\}$. \mathcal{N} denotes the D-dimensional Gaussian distribution:

$$\mathcal{N}(X|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right). \tag{4.2}$$

Expectation maximization $(EM)^4$ [DLR77] is usually used to solve a GMM by the following steps:

E-step: estimate the posterior probability p_{ij}^t at t iteration as:

$$p_{ik}^{t} = \frac{\pi_{k}^{t} p(x_{i} | \mu_{k}^{t}, \Sigma_{k}^{t})}{\sum_{k=1}^{K} \pi_{k}^{t} p(x_{i} | \mu_{k}^{t}, \Sigma_{k}^{t})}$$
(4.3)

M-step: update the parameters μ_k , Σ_k and π_k at t+1 iteration based on the probabilities from the E-step:

⁴ Since EM has great impact on machine learning and data mining applications, we give the theoretical derivation of EM in Appendix A.3.

$$\pi_k^{t+1} = \frac{1}{N} \sum_{i=1}^{N} p_{ik}^t, \tag{4.4}$$

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{N} p_{ik}^t x_i}{\sum_{i=1}^{N} p_{ik}^t},\tag{4.5}$$

$$\Sigma_k^{t+1} = \frac{\sum_{k=1}^K p_{ik}^t (x_i - \mu_k^t) (x_i - \mu_k^t)^T}{\sum_{i=1}^N p_{ik}^t}.$$
 (4.6)

EM is guaranteed to converge, so that a locally optimal solution is always assured. However, one big concern in GMMs is the number of components/clusters (K) that must be defined in advance, which can be a hard task. We employ the Bayesian information criterion (BIC) [Sch78] to determine this parameter automatically. The BIC is frequently used for model selection, since it considers a trade-off between model fitting and model complexity. By adding a penalty term for the number of parameters in the model, the BIC can alleviate the problem of overfitting, which can be caused by increasing the likelihood by just adding parameters to the model. The BIC has the form:

$$BIC = -2\log(\text{likelihood}) + \log(N)P, \tag{4.7}$$

where P is the number of free parameters to be estimated in the model and N is the total number of data points. In a multivariate Gaussian, the number of free parameter is (KD(D+3))/2 + K - 1, due to K-1 mixing proportions to decide, KD mean values, and (KD(D+1))/2 free parameters in the covariance matrix. The model with the lowest value of BIC is selected as the desired model. Aside from BIC, the Akaike information criterion (AIC) [Aka74] is also a common method for model selection.

$$AIC = -2\log(likelihood) + 2P,$$
(4.8)

The log-likelihood (\mathcal{L}) of the GMM can be inferred as:

$$\mathcal{L}(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k) \right\}$$
(4.9)

$$\begin{split} &= \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} A_k \right\}, \ A_k \text{ is a multivariate Gaussian } \mathcal{N}(X|\mu, \Sigma) \\ &= \sum_{n=1}^{N} \left\{ \log A_m + \log \left(\sum_{k=1}^{K} \exp(\log A_k - \log A_m) \right) \right\}, \text{ log-sum-exp trick} \\ &\text{let } A_m = \max \{ \pi_i \mathcal{N}(x_n | \mu_i, \Sigma_i) \}, i \in 1, \cdots, K \\ &= \sum_{n=1}^{N} \left\{ \log \pi_m - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_m| \right\} - \\ &\sum_{n=1}^{N} \left\{ \frac{1}{2} \sum_{n=1}^{N} (x_n^m - \mu_m)^T \Sigma_m^{-1} (x_n^m - \mu_m) \right\} + \\ &\sum_{n=1}^{N} \left\{ \log \left(\sum_{k=1}^{K} \exp(\log A_k - \log A_m) \right) \right\}, \end{split}$$

where the "log-sum-exp" trick [Mur12] is a method for avoiding numeric underflow and thus can improve the numeric stability when computing the BIC in a GMM scenario. Continue writing out the Eq. 4.9, we can compute the log-likelihood of the GMM, which is needed for computing Eq. 4.7 and Eq. 4.7.

$$\log\left(\sum_{n=1}^{N} A_n\right) = \log A_m + \log\left(\sum_{n=1}^{N} \exp(\log A_n - \log A_m)\right),\tag{4.10}$$

where A_m is the largest term among A_n . Continue writing out the Eq. 4.9, we can compute the log-likelihood of the GMM, which is needed for computing Eq. 4.7 and Eq. 4.8.

It can be seen that AIC (the lower the better) has a lower penalty for model complexity, because $\log(N)P$ usually is much larger than 2P. Because both AIC and BIC consist of two terms, the final scores weight the relevance of these two terms. In AIC, the term 2P does not contribute much to the final score as opposed to $-2\log(\text{likelihood})$. However, $\log(N)P$ increases much faster when the model becomes more complex (more clusters), thus the resulting BIC score stops growing as the number of clusters increases. From Eq. 4.7 and 4.8, we see that AIC and BIC suggest different quantities. Thus, they may yield different model selection results. In the present study, we conduct experiments using both AIC and BIC, and report more results on BIC. In the following, we review some reported studies comparing BIC with AIC. Our experiments show that AIC and BIC yield comparable results,

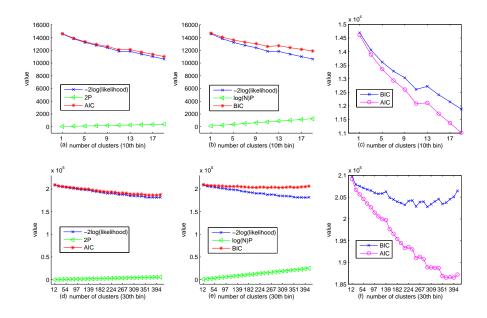


Figure 4.1: BIC and AIC score displayed for the 10th (top) and 30th (bottom) bin in 50 bin using pSMC normalization of the ADNI dataset. The AIC (red) and BIC (red) values are simply $-2 \log(\text{likelihood})$ (blue) plus 2P (green) and $\log(N)P$ (green), respectively. Refer to Eq. 4.8 and Eq. 4.7 for the definition of AIC and BIC. (a), (b) (d) and (f) show the AIC and BIC along with their components' value. (c) and (f) offers a direct comparison between AIC and BIC. The 10th bin contains 586 voxels, hence it is a small sample, in contrast to the large sample of the 30th bin containing 12,414 voxels. Note that the term "sample" in this context does not mean the number of patients, but the number of voxels to be clustered.

Table 4.2: Grade of evidence of the BIC difference [KR95].

BIC difference	Evidence
0-2	weak
2-6	positive
6-10	strong
>10	very strong

cf. Tables 4.3 and 4.4.

A recent study [DBU⁺09] reveals a similar observation that illustrates model selection discrepancy between AIC and BIC, preferring BIC to AIC for the sake of favoring simpler models. In addition, another work [PZKM13] explicitly states that "AIC has been shown to perform well at selecting the true number of factors when

it exists, but only at small sample size N. BIC has been found to outperform AIC in recovering the true K (number of clusters)". Particularly, we observe that AIC and BIC tend to agree with each other when the number of voxels (sample size) is small (i.e., 586, cf. Fig. 4.1). As the number of voxels increases to 12,414, BIC reaches the lowest value much earlier than AIC, which is illustrated in Fig. 4.1. Thus, BIC is more appropriate when the number of voxels is large (greater than 3500 may mean "large", refer to work [BA04] to avoid a too complex model. As for the plot (c) in Fig. 4.1, it can be seen that the AIC and BIC continues dropping as the number of clusters increases. Thus, one may conclude that the more clusters, the better in the small sample case. However, we can, at least, observe a dip if we test a larger number of clusters. Because we intentionally impose that a cluster should at least contain 30 voxels to avoid trivial clusters, more clusters are not necessary to be tested (refer to Section 4.1.3.1 Parameters in Model).

Another study shows that BIC performs better than AIC in both small and large sample size cases, claiming that AIC lacks the appropriate penalty to prevent over-fitting [NST07]. The difference between our BIC models is greater than 10 (see plot (c) and (f) in Fig. 4.1), thus it suggests strong evidence (cf. Table 4.2) that the model difference is meaningful according to work on Bayes factors [KR95].

To empirically show the difference between AIC and BIC in this study, we show the generalization error (GE, equivalently, 1–accuracy) demonstrated by Fig. 4.2. From the 50 bins to 150 bins, we observe that BIC outperforms AIC on 8 cases (50, 60, 70, 80, 90, 110, 120, 140 bins) in terms of mean generalization error. BIC often yields a certain bin that has the lowest GE, for example, in 60 bins, 70 bins, 90 bins, 110 bins, 120 bins, 140 bins and 150 bins. When we perform the classification using all features extracted on all bins together (details are explained in Section 4.1.2.3 GMM+MS on 3D PET images), we see that BIC shows lower GE than AIC demonstrated by the bottom plot of Fig. 4.2. The reason for this is that selected features (on all bins) by BIC contain more discriminative information than the features by AIC, which complies with the results of MCI against AD using pSMC normalization using ADNI dataset. However, AIC shows better performance than BIC in some other cases, such as MCI against AD using the TUM dataset. In summary, AIC and BIC suggest generally comparable results across the two datasets in different classification cases.

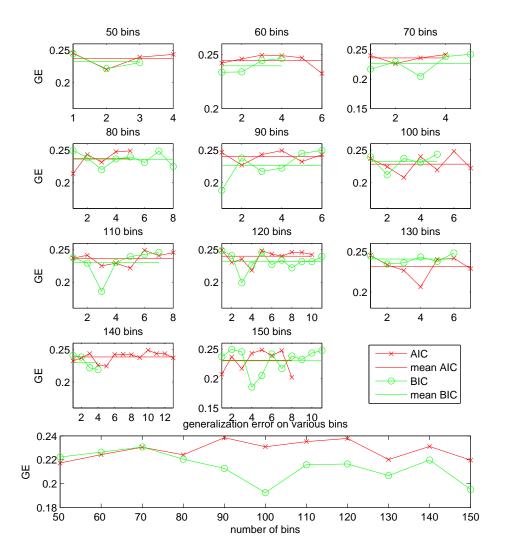


Figure 4.2: Generalization error (GE) by BIC and AIC using pSMC normalization of the ADNI dataset. The generalization error (equivalently, 1–accuracy) is computed based on features (mean and standard deviation value) extracted from the clusters in the individual bin (one bin among 50 bins, for example) via 10 times 10-fold cross-validation. For example, a green point can represent the classification GE using the n-th bin when we divide the whole brain voxels into 50 bins. The plotted points are the bins that yield GE < 0.25, and these bins are highly predictive, thus contribute to the final classification results. The x-axis represents only the number of points, implying no ordering. The horizontal green line and red line are the mean values computed from green circles and red crosses, respectively. The bottom plot illustrates the generalization error from dividing brain voxels into 50 bins till 150 bins after collecting top informative features from the individual bins.

4.1.2.3 GMM+MS on 3D PET Images

We employ a clustering method (GMM) to group brain voxels into small regions that exhibit both high similar intensity and geometric affinity. A PET image can be viewed as three dimensional (3D) spatial data along with one extra dimension that represents the intensity of each voxel. Thus, a voxel is denoted by a 4-tuple $(x,y,z,I) \in \mathbb{R}^4$, where x,y,z are the spatial coordinates and I is the intensity value. We used NC PET images as reference images to obtain the clustering results and used these clusters to extract the features from the NC, MCI and AD groups. Note that the method is applied on the AAL (gray matter voxels of MNI space) defined voxels, which constitute the gray matter in the brain. The mean intensity and standard deviation of each cluster are subsequently defined as features. To ensure the clusters have similar intensity values and are geometrically connected, we first group the original voxels into a certain number (e.g., 100) of bins of equally sized intensity ranges, and only then cluster each bin by the introduced methods based only on the spatial information, i.e., the x, y, z coordinates. A bin is a data interval that is described by a statistical histogram. The data falling into the same bin are from a certain value interval, such that the data within the same bin are similar in their values. Theoretically, it is hard to find the most appropriate number of bins in advance, thus we tested different numbers from 50 to 150 with step size 10, i.e., 50, 60, ..., 150. The best one can be chosen by a cross-validation on the training data. In practice, we can keep track of the same cross-validation sampling, so that the training data are ensured to be identical for running the various bins. Given the training data, we can further split them into sub-training and sub-test data, which are used to train and evaluate the model respectively. Evaluating the model using the sub-test data gives a predictive accuracy. The yielded highest accuracy of a certain bin corresponds to the most appropriate number of bins. Once the desired number of bins is found, the same training procedure can be applied to the whole training data to maximize the use of present training data. Therefore only the training data is used to set the optimal parameters in the experiments. The workflow of the proposed method is summarized in Algorithm 4.1 as well as the Section 4.1.2.6.

If there are 1000 clusters formed using GMM+MS, then the image can be repres-

Algorithm 4.1: Workflow of proposed GMM+MS clustering method on 3D PET images.

- Stratified 10-fold cross-validation.
- Training phase (9 folds) Run different bins 50, 60, 70, ..., 150. Split training data into disjoint sub-training and sub-test data.
 - 1. Given a Mean NC image, divide the AAL defined voxels into the specified number of bins, such as 50. Mean NC is averaged over all NC images.
 - 2. For each bin, run the GMM+MS method to yield the clusters.
 - 3. Collect all the resulting clusters from all the bins.
 - 4. Given a NC, MCI or AD image from the pool of sub-training data, compute the mean (μ) and standard deviation (σ) of the voxels in each cluster using the provided spatial information, i.e., of the cluster.
 - 5. The image can then be represented as a feature vector by the mean (μ) and standard deviation (σ) .
 - 6. Build SVM model using only the sub-training data, and the predictive accuracy is computed for the sub-test data using the model. (The model is trained on MCI and AD sub-training data, if the classification is MCI against AD.)

Collect the computed accuracy from all bins.

- The resulting clusters correspond to the bin with the highest accuracy are used as the most appropriate clusters for NC, MCI and AD images.
- Test phase (remaining 1 fold)
 - 1. Construct the features for both training and test images described as the steps of 4 and 5.
 - 2. Build SVM model using the training data (9 folds), and obtain the results using the remaining test data (1 fold).

ented as a 2000 (1000 μ and 1000 σ) dimensional vector. Generally, not all of these features are informative, thus we applied a feature selection technique [CL06] to choose the most discriminative ones for building the model. In this study, we empirically used the top-150 most informative features for learning the model. From Fig. 4.3 of BIC, we observe that the classification accuracy increases in the beginning, but

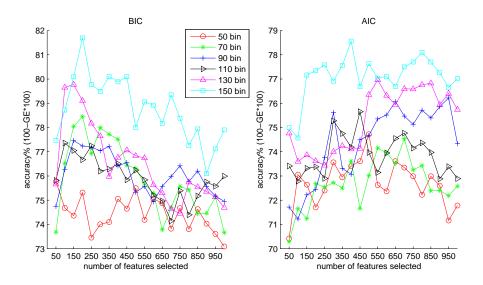


Figure 4.3: Number of features selected versus the classification accuracy using pSMC normalization of the ADNI dataset on MCI against AD. Accuracy of BIC tend to drop after selecting 150 features, and AIC tends to drop after 400.

drops after selecting too many features. 150 features appear to provide sufficient classification relevant information, and more features may hamper the classifier's performance due to the well-known curse of dimensionality [Bis06]. As for AIC, top-400 features were selected to perform the experimental comparison depicted by Fig. 4.3. AIC needs more features than BIC, which may be the reason why AIC divides voxels into more clusters so that the discriminative information are spread over many clusters. It should be pointed out that the feature selection [SFK10] and the model building steps only used information from the training set: no information from the test set is used at any point in time (in other words, no information leakage from the test set to the training set has occurred).

A support vector machine (SVM) was used to build the final classification model, which is trained on the training data. The SVM has been shown to perform well in a variety of applications, thus it was chosen to be the classifier in this study. A tutorial [Bur98] offers a good introduction to the SVM. Apart from the SVM, other classification methods could be used, such as Random Forests, Naïve Bayes, and others. We do not attempt to compare the proposed method with SVMs, we rather use an SVM as a classifier in the method. The suggested method aims at extracting useful features from brain voxels, whereas SVMs are a classification method based

on input features (voxels).

In terms of running time, it costs roughly 15 hours to cluster the mean NC image from 50 bins to 150 bins. It takes only a few seconds to extract the features of a given PET scan after having the clusters. Therefore, the proposed approach is very efficient once the clusters are derived, since extracting features from future images is fast. The code is implemented in MATLAB and runs on a machine with Intel(R) Core(TM) i7-3632QM CPU @2.20 GHz, 8GB of memory. In addition, the LIBSVM [CL11] package provides a fast classification, once the features are constructed.

4.1.2.4 Compared Methods

AAL approach: AAL [TMLP⁺02] defines 116 brain regions, and we extracted the mean and standard deviation from each of these regions as features. Thus, in total each image is represented by a 232-dimensional feature vector.

t-test: This hypothesis testing based method uses voxel-wise t-test and is widely applied in neuroscience studies. A t-test based method, for example, was tested as a method for feature selection with respect to predictive accuracy recently [CHC⁺12]. If the p-value is lower than a pre-defined threshold, e.g., 0.001, then this voxel is regarded as an indicator voxel for two groups of individuals. The null hypothesis is that the voxels in the two groups come from a population where the means of the two groups are the same. Therefore, a p-value lower than 0.001 rejects the null hypothesis, i.e., the two groups have different means. Hence, this voxel can be an indicator voxel representing group difference. We performed a two-tailed t-test on each of the voxels defined in an AAL region without multiple comparison correction, with a threshold set to 0.001. Finally, the top-150 voxels with the lowest p-value were chosen for learning, to be in line with our proposed method.

4.1.2.5 Test Protocol

The experiments were performed on two datasets (ADNI and TUM, cf. Table 4.1) independently. For each of the datasets, we trained the model based on the training data, evaluating the model using the held-out test data. The training and test data were divided using 10-fold cross-validation. In statistics, 10-fold cross-validation is commonly applied as an approach to testing a predictive model. In 10-fold cross-validation, technically, the original dataset is roughly divided into ten subsets, and

each time we select nine subsets for training the model and the remaining one for testing. This procedure is repeated ten times, assuring that each subset is tested exactly once and employed nine times for training. The ten results are then averaged to produce a single estimate. (In case of very small data sets, results are aggregated by instance and not by test set.) In the two-class classification case, each subset contains roughly the same proportion of samples from the two classes, which is called *stratification*. For MCI versus AD in the TUM dataset (30 MCI, 30 AD), for instance, each subset consists of six samples with three from the MCI and three from the AD group. Each time 54 samples are employed for training and six for testing. After repeating the procedure ten times, we compute the mean value from all ten runs as the final result. We used the implementation of "crossvalind" function in MATLAB 2010 (R2010a) to achieve the stratified 10-fold cross-validation.

4.1.2.6 GMM+MS in Summary

To sum up, we tested different numbers of bins (50, 60 to 150 with a step size of 10) to decide how many bins the whole brain voxels should be divided into. For example, if the brain voxels are divided into 50 bins, we run the GMM algorithm on each of the 50 bins. The BIC suggests the optimal number of clusters in each bin. The yielded clusters from the 50 bins are then collected as the final clusters in the whole brain. The mean and standard deviation of each cluster (given x, y, and z coordinates) are extracted as the feature values representing a PET scan. Finally, every PET scan can be represented by a vector. A 10-fold cross-validation is then used to train an SVM and test the classification performance based on the vectors. Algorithm 4.1 further explains the procedure of the method applied to PET scans.

4.1.3 Experiments

4.1.3.1 Parameters in Model

The proposed method is able to find the optimal number of clusters by comparing the BIC score computed from different models. Thus, a number of different model selections must be performed. The simplest way to determine the number of clusters in a model is to let the cluster number be equal to the number of voxels in each studied bin (a voxel value interval). However, this is not only computationally expensive, but we may also end up with clusters that include too many (too rough) or too few (too trivial) voxels. To this end, we intentionally defined a priori that the resulting clusters should have 1000 voxels at most and 30 voxels at least. Then the maximal number of clusters in a bin can be calculated as Cmax=#(voxels in a bin)/1000, and the minimal is Cmin=#(voxels in a bin)/30. The numbers between the Cmax and Cmin are used in these model selections. Fig. 4.4 depicts that the number of derived clusters, in general, decreases as the number of bins increases. However, the sharp drop appears in the beginning and the curve then gradually reaches a nearly stable state until 150.

4.1.3.2 Performance Comparison

Table 4.3 shows that the GMM+MS method performs much better than the AAL and t-test methods on the task of discriminating MCI from AD. In particular, GMM+MS is statistically significantly better than the t-test method. Specifically, regarding pSMC, the accuracy gain is 12.9% (80.2%-67.3%) with a p-value of 0.017 (calculated using the test suggested by [BF04]), and the specificity gain is 0.19 (0.80-0.61) with a borderline p-value of 0.066. Regarding the grand mean normalization method, the accuracy, AUC, sensitivity and specificity all show better results than the AAL and t-test approaches. As for NC versus AD, the three methods perform equally well, which may due to the fact that the most essential discriminative information can be easily identified by all of them. As a result, further improvement hardly can be achieved. The t-test approach reveals a slightly better result on NC versus MCI using pSMC, whereas it shows similar performance using grand mean normalization. However, the opposite is true on the TUM dataset (cf. Table 4.4), which suggests that GMM+MS is much better than the other two methods. Again, a comparable performance is shown for NC versus AD. Regarding MCI versus AD, the grand mean still reveals improved results, and pSMC shows comparable performance. As for the comparison between AIC and BIC, there is no significant difference revealed by the results, cf. Tables 4.3 and 4.4. A possible explanation is that both AIC and BIC can discover discriminative clusters sufficiently well, despite the fact that AIC favors a more complex model and BIC tends to choose a simpler one.

The ROC (receiver operating characteristic) curves shown in Fig. 4.5 and 4.6 reveal the different performance of various methods, depicting the false positive rate against the true positive rate. The BIC (green) and AIC (red) curves cover a large portion of the t-test and AAL curves regarding MCI versus AD on both ADNI and

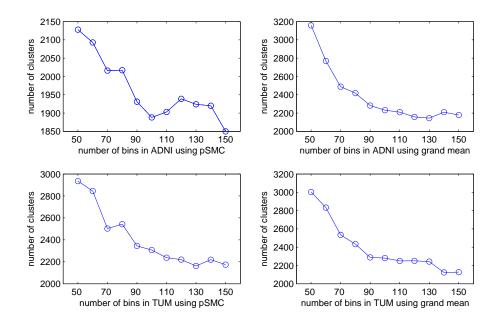


Figure 4.4: Relation between the number of clusters and the number of bins on ADNI and TUM datasets based on the normal control PET scans.

TUM datasets. This observation complies with the results shown in Tables 4.3 and 4.4, i.e., the proposed method GMM+MS performs better than the compared ones in terms of MCI versus AD.

To sum up, the proposed method performs substantially better than the compared methods, in particular for MCI versus AD. Specifically, three comparisons out of four (TUM and ADNI datasets, grand mean and pSMC methods) demonstrate improved performance. A statistically significant result is also confirmed on the ADNI dataset. The limited NC sample size of 16 in the TUM dataset, as opposed to 30 in ADNI, may be one reason for less accurate results.

4.1.3.3 Results Analysis

In general, the proposed method achieved either salient performance improvement or comparable results compared to more established methods using two different normalization methods on two independent datasets. However, the results are not perfectly consistent across the two datasets, which might be caused by different types of scanners and different image acquisition methods, such as the amount of

Table 4.3: Result summary of three different methods on ADNI dataset. "*" denotes the GMM+MS is significantly better than t-test approach at a statistical level of 0.05. The p-value was calculated by the corrected paired t-test tailored for comparing learning algorithms [BF04]. A library for support vector machines (LIBSVM) [CL11] was used to build the SVM models. A linear kernel is used, with a grid search for parameter optimization. Grid search considers only the optimization of the penalty parameter C in the linear SVM, selecting the value of C yielding the best classification result based on the training data. After the best value of C is found, we apply it to the test data. AUC: area under ROC curve. Each experiment was repeated 10 times with a 10-fold cross-validation. P: results using "primary sensorimotor cortex" region for intensity normalization. G: results using "grand mean" method for intensity normalization. Results from AIC are noted in brackets, following the BIC results.

			Accuracy%	AUC	Sensitivity	Specificity
	Р	GMM+MS	80.2* (78.3)	$0.85\ (0.82)$	0.80 (0.77)	0.80 (0.79)
		AAL	74.2	0.81	0.75	0.74
MCI vs. AD		t-test	67.3	0.79	0.72	0.61
MCI vs. AD	G	$_{\mathrm{GMM+MS}}$	$77.1\ (78.1)$	$0.83\ (0.83)$	$0.85\ (0.78)$	$0.68 \; (0.77)$
		AAL	73.2	0.80	0.77	0.68
		t-test	69.5	0.81	0.76	0.62
	Р	GMM+MS	89.1 (88.4)	0.97 (0.97)	0.92 (0.91)	0.86 (0.85)
		AAL	88.2	0.97	0.90	0.86
NC vs. AD	G	t-test	89.1	0.97	0.92	0.85
NC vs. AD		$_{\mathrm{GMM+MS}}$	87.7 (88.1)	$0.96 \ (0.97)$	0.93 (0.91)	0.81 (0.83)
		AAL	88.8	0.96	0.90	0.87
		t-test	87.1	0.95	0.93	0.79
	Р	GMM+MS	63.2 (62.9)	0.72 (0.72)	0.65 (0.67)	0.61 (0.59)
		AAL	63.7	0.81	0.66	0.60
NC vs. MCI		t-test	67.1	0.79	0.68	0.65
NO vs. MOI	G	$_{\mathrm{GMM+MS}}$	64.6 (61.3)	0.74 (0.72)	$0.66 \ (0.66)$	$0.64 \ (0.56)$
		AAL	63.7	0.80	0.67	0.60
		t-test	65.8	0.81	0.66	0.65

tracer used, whether an eye mask was used during the scan, and so forth.

The information in Tables 4.5 and 4.6 highlights the detailed brain region information regarding the contribution to the classification. These regions include areas which are typically involved in AD, such as the cingulum, precuneus and temporal regions. The red points in Fig. 4.7 and 4.8 highlight these informative voxels (brain regions), which correspond to the information in Tables 4.5 and 4.6.

The use of two independent datasets is a major strength of our study. In addition,

			Accuracy%	AUC	Sensitivity	Specificity
	Р	GMM+MS	72.7 (74.0)	0.81 (0.80)	0.77 (0.78)	0.68 (0.69)
		AAL	74.8	0.80	0.77	0.72
MCI vs. AD		t-test	72.6	0.79	0.77	0.68
MCI vs. AD	G	$_{\mathrm{GMM+MS}}$	$73.8 \ (74.8)$	$0.82\ (0.82)$	$0.77\ (0.78)$	$0.71\ (0.71)$
		AAL	70.5	0.78	0.72	0.69
		t-test	65.5	0.71	0.68	0.63
	Р	GMM+MS	90.5 (89.5)	0.93 (0.96)	0.94 (0.88)	0.89 (0.89)
		AAL	89.0	0.95	0.94	0.85
NC vs. AD		t-test	89.4	0.97	0.86	0.90
NC vs. AD	G	$_{\mathrm{GMM+MS}}$	91.6 (88.1)	0.97 (0.96)	0.94 (0.84)	0.90(0.89)
		AAL	89.0	0.95	0.91	0.88
		t-test	92.0	0.98	0.93	0.91
	Р	GMM+MS	88.0 (87.2)	0.95 (0.93)	0.87 (0.85)	0.89 (0.89)
		AAL	81.5	0.88	0.82	0.79
NO MOI		t-test	81.1	0.90	0.64	0.89
NC vs. MCI	G	$_{\mathrm{GMM+MS}}$	87.7 (85.3)	0.95 (0.92)	0.87(0.82)	0.88(0.87)

Table 4.4: Result summary of three different methods on TUM dataset. Results from AIC are noted in brackets, following the BIC results.

two different intensity normalization methods, namely primary sensorimotor cortex and grand mean normalization, were applied to establish the experimental results. Tables 4.3 and 4.4 show that the proposed method is better compared to two other methods.

0.90

0.94

0.74

0.89

0.80

0.90

78.0

89.5

4.1.3.4 Error Analysis

AAL

t-test

To gain some insights into the classification difference between GMM+MS, AAL and the t-test methods, we exert to investigating the errors committed by the classifier. To be concise and illustrative, we take MCI against AD (grand mean normalization) in the ADNI dataset as a running example. Table 4.3 shows that the proposed method leads to an accuracy gain of 7.6% (77.1%-69.5%), which is a salient improvement. Looking at the misclassified images, we observe that a certain image is misclassified by AAL 4 times and 10 times by the t-test approach, but without any misclassification by GMM+MS. Therefore, this image, in fact it is an AD image,

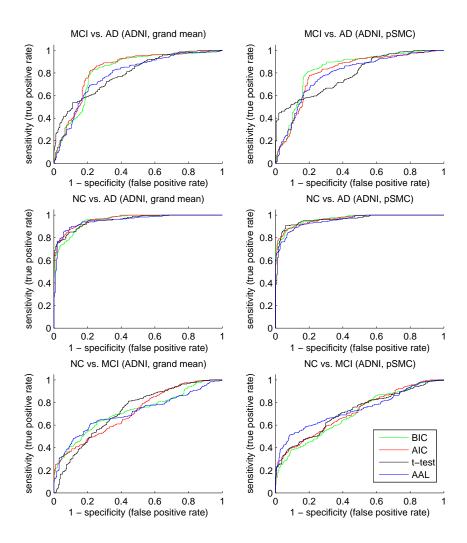


Figure 4.5: ROC curve of compared method on ADNI dataset. To plot the curve, we collected the predicted probabilities for all the test sets in 10 times 10-fold cross-validation, along with their true class labels. Refer to page 173 in the book written by Witten *et al.* [WFH11] for the ROC curve plotting.

is selected for a more detailed study. Recall that the SVM needs to compute the sign of y = wx + b to make a decision. Hence, knowing w and b is essential. Since b is only a constant, we omit it from further analysis. The studied AD image is regarded as the test data and all the rest is treated as training data. After training

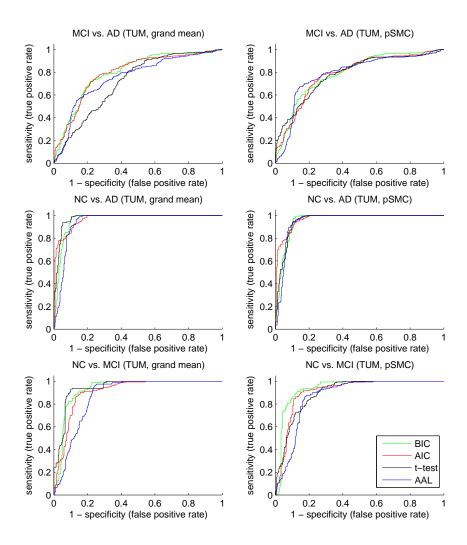


Figure 4.6: ROC curve of compared method on TUM dataset.

the model, the SVM returns the support vectors and the weights w (feature importance), such that the classification can be made upon y = wx + b. To stay illustrative and straightforward, we take a closer look at the Euclidean distance, since it gives a direct impression of dissimilarity. To this end, all the data, training and test, are multiplied by the weight vector w, such that each instance is re-weighted by their importance in terms of the SVM. Subsequently, the Euclidean distance is calculated

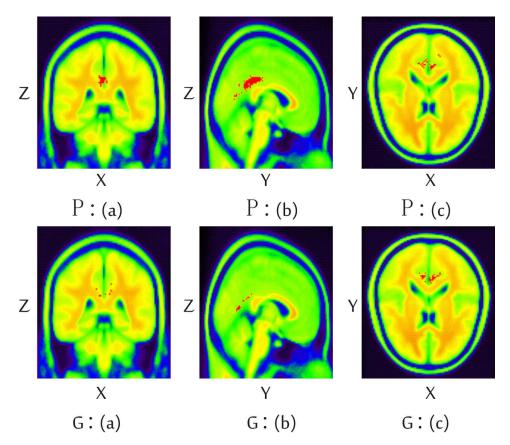


Figure 4.7: The informative regions (voxels) of MCI against AD using ADNI dataset of the 45th layer. (a): coronal view (b): sagittal view (c): transaxial view. P: primary sensorimotor cortex normalization; G: grand mean normalization. x, y and z are the width (91), depth (109) and height (91) respectively. The red points represent the informative voxels, whereas other colors are only used to depict the brain structure.

between the test image and each training image (MCI and AD group). Finally, the mean weighted Euclidean distance is computed for the MCI and AD group, respectively, which represents the dissimilarity between the test image and the group. We compute the relative ratio to denote the dissimilarity:

$$\rho = \frac{D_{\text{MCI}} - D_{\text{AD}}}{D_{\text{AD}}},\tag{4.11}$$

where D_{MCI} is the mean distance between the test image and the MCI group, and the same as for D_{AD} . The greater the ρ , the more similar to AD. As a result,

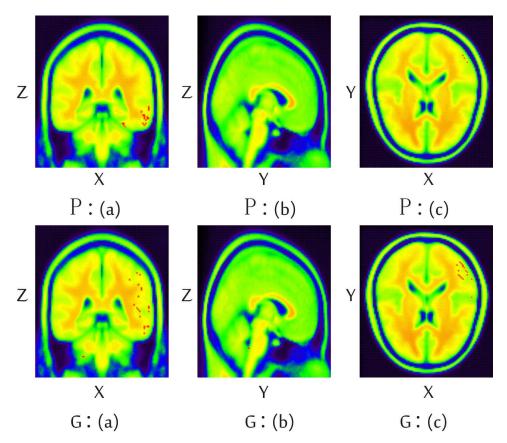


Figure 4.8: The informative regions (voxels) of MCI against AD using TUM dataset of the 45th layer. (a): coronal view (b): sagittal view (c): transaxial view. P: primary sensorimotor cortex normalization; G: grand mean normalization. x, y and z are the width (91), depth (109) and height (91) respectively. The red points represent the informative voxels, whereas other colors are only used to depict the brain structure.

 $\rho(\text{GMM+MS}) = 0.26$, $\rho(\text{AAL}) = 0.07$, $\rho(t\text{-test}) = 0.03$. Hence, GMM+MS indicates the greatest value and thus it classifies this image correctly as AD. Therefore, the features derived from GMM+MS enable the SVM to make the correct decision in this case, in contrast to the AAL and t-test based methods.

4.1.4 Discussion

In this work, a machine learning approach, GMM+MS, is used to derive clusters based on an averaged NC PET image. The proposed method has the advantage of determining the number of clusters automatically, using a widely accepted model

Table 4.5: Informative regions (voxels) of MCI against AD using ADNI dataset. The top-ten clusters in GMM are recorded in each cross-validation, with different scores assigned to these clusters. The most informative cluster has the highest score. After 10 times 10-fold cross-validation, we rank these clusters according the overall score and select the top-ten, which are marked by the red points in Fig. 4.7. Within these ten clusters, their corresponding AAL brain regions are identified and ranked according to the proportion denoted by the numbers. The region names remain as noted in the AAL template. The proportion is calculated as the number of significant voxels in the region divided by the number of total significant voxels.

1	P: Cingulum_Post_L:	18.4%	6	P: Cingulum_Mid_L:	5.74%
	G: Precuneus_L:	28.4%		G: Cingulum_Mid_L:	5.28%
2	P: Precuneus_L:	20.4%	7	P: Cerebellum_8_L:	2.87%
	G: Cingulum_Post_L:	21.1%		G: Calcarine_R:	5.05%
3	P: Precuneus_R:	17.9%	8	P: Cuneus_L:	3.53%
	G: Precuneus_R:	14.2%		G: Cingulum_Mid_R:	4.36%
4	P: Cingulum_Mid_R:	15.7%	9	P: Calcarine_R:	2.06%
	G: Cingulum_Post_R:	10.0%		G: Cuneus_L:	3.90%
5	P: Cingulum_Post_R:	10.0%	10	P: Occipital_Sup_L:	1.33%
	G: Calcarine_L	5.50%		G: Lingual_R:	1.15%

Table 4.6: Informative regions (voxels) of MCI against AD using TUM dataset.

		()		
1	P: Temporal_Mid_L:	29.3%	6	P: Cerebelum_6_L:	3.41%
	G: Temporal_Mid_L:	40.8%		G: Parietal_Inf_L:	3.75%
2	P: Temporal_Inf_L:	29.0%	7	P: Parietal_Sup_L:	1.28%
	G: Temporal_Inf_L:	21.5%		G: Postcentral_L:	3.65%
3	P: Fusiform_L:	13.0%	8	P: Angular_L:	1.28%
	G: Occipital_Inf_L:	8.74%		G: Fusiform_L:	2.11%
4	P: Occipital_Inf_L:	10.8%	9	P: Cerebelum_Crus1_L:	1.19%
	G: Occipital_Mid_L:	8.65%		G: Temporal_Sup_L:	1.83%
5	P: Occipital_Mid_L:	6.47%	10	P: Precuneus_L:	1.11%
	G: SupraMarginal_L:	4.42%		G: Angular_L:	1.25%

selection criterion. The model selection procedure assures that the derived model has a good trade-off between model fitting and model complexity. In such a way, a too complex model can be excluded, although it may have a good degree of model fitting. On the other hand, if a model is too concise, it may not have a satisfying level of model fitting. Therefore, the model selection procedure aims to keep the model complexity in a good balance. The two-phased algorithm first divides the

voxel values into different bins and then applies the GMM+MS on the coordinate information at each of the bins to yield the final clusters. The resulting clusters have similar intensity values and are also geometrically connected. The experimental results suggest that the proposed method can outperform the compared methods, in particular for discriminating MCI from AD. The underlying reason can be that the proposed algorithm is able to discover finer (smaller) clusters that are helpful in discriminating MCI from AD, while the AAL and t-test approach may fail to reveal such critical information. However, a little inconsistency is seen by the different intensity normalization methods, which also suggests that the intensity normalization procedure can be an important factor. In the previous section, we also try to shed some light on the performance difference between these methods. Although the SVM is usually applied as a black-box classifier, we can still employ the support vectors and the weights to gain important insights. Since there are 150 features for the GMM+MS and t-test methods, and 232 for AAL, they are high-dimensional datasets, which makes it hard to analyze which features contribute to the correct classification in the end. However, by introducing the relative ratio computed from the Euclidean distance, it is possible to quantitatively show the difference between these approaches. In terms of time complexity, the AAL method is the fastest because it is based on pre-defined brain regions. GMM+MS needs to work further based on defined AAL regions. In addition, the number of bins tested can also influence the running time. As for the t-test method, it is simple, but requires more memory to store the images for a group comparison, which can sometimes become a problem if there are too many images. The proposed algorithm can be widely applied as a feature extraction method on medical imaging data, which can assist the medical imaging community to discover interesting discriminative brain voxels pattern. The applicability of the algorithm may reach broader application scenarios than merely AD classification, as long as imaging feature extraction is concerned. In particular, we also provide a thorough study on the comparison between AIC and BIC, which offers a clear guidance for the model selection issue. One limitation of the work is the open problem of discriminating patients with MCI who progress to clinically diagnosable AD from those who remain clinically stable: this remains an important and challenging task. To deal with it, one may need a clearly defined dataset (MCI follow-up) and a reasonably sound algorithm, which is left for future work.

4.1.5 Conclusions

The present work proposed a new clustering method, i.e., GMM+MS, for FDG-PET images. It has the advantage of determining the number of clusters by a model selection approach. This method is applied only on an NC image to define the clusters, and then the resulting clusters can be used to extract features from the MCI and AD images. Throughout the experiments on two independent datasets, we not only demonstrate the merits of suggested method, but also show that the intensity normalization and different datasets (acquired by different scanners) indeed play some role in the results. In conclusion, our results suggest that the discriminative information in the FDG-PET images can be extracted by the proposed approach.

4.2 Survival Analysis: A Follow-up Case Study

4.2.1 Motivation

In Section 4.1.2.2, we proposed method for classification of different dementia forms, such as MCI versus AD. In this section, we investigate the differentiation between patients who remain in the MCI stage (MCI_{MCI}) and those that progress to AD dementia (MCI_{AD}). The classification between MCI_{MCI} and MCI_{AD} is of great medical interest, because it may reveal valuable knowledge in dementia studies. To this end, we first apply survival analysis and infinite Gaussian mixture model to FDG-PET scans in order to distinguish MCI_{MCI} from MCI_{AD}. Second, we study the usefulness of non-imaging data compared to FDG-PET imaging data.

4.2.2 Materials and Methods

4.2.2.1 Study Participants

Experiments were performed using data from the publicly available AD Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/) accessed in the year 2011. Only data from the first stage of ADNI (ADNI 1) were considered. ADNI has a large pool of FDG-PET (co-registered, averaged) images, which have been acquired on various scanners using different imaging parameters. To eliminate bias due to these factors, we selected images that had been obtained using the same scanner (Siemens/CTI) as well as the same parameters, such as the number of slices.

To ensure that differences in the degree of cognitive impairment at baseline did not affect our results, we only included patients with a Clinical Dementia Rating (CDR) sum of the boxes score of 1.5, 2 and 2.5 [Mor93]. Patients were only considered if they had at least five consecutive clinical assessments every six months up to a maximum of 24 months in order to have sufficient data for the survival analysis. The planned pre-selection of participants resulted in a total of 77 patients, including 45 MCI_{MCI} and 32 MCI_{AD}. MCI_{MCI} was defined as patients not meeting the Institute of Neurological and Communicative Disorders and Stroke-AD and Related Disorders Association (NINCDS-ADRDA) criteria for AD at their last follow-up assessment, whereas MCI_{AD} patients met the NINCDS-ADRDA criteria at least at one of the follow-up visits. It is important to note that this is a censored dataset, which includes patients that did not progress to AD dementia until their last follow-up visit for various reasons including withdrawal of consent, death or limited length of follow-up. This phenomenon is often studied using survival analysis [KK11], which is also the essential technique applied in this work. The characteristics of the study population are summarized in Table 4.7.

Time to progression from MCI to AD dementia was calculated as the time between the baseline visit and the visit at which an AD dementia diagnosis was first established. As a result, we obtain a dataset that allows us to train a model predicting progression to AD dementia within 24 months.

4.2.2.2 Image Pre-processing

Prior to their use for image analysis, the FDG-PET images underwent two preprocessing steps in the statistical parametric mapping software package SPM5 (Wellcome Functional Imaging Laboratory, London, UK), based on Matlab R2010a (The Mathworks Inc, Natick, USA): spatial normalization and smoothing (kernel size [8 8 8] mm). Spatial normalization ensures that the processed image is of the size 91×109×91 voxels, which is in accordance with the Anatomical Automatic Labeling (AAL) template [TMLP+02]. The final step is intensity normalization, which was done by dividing each voxel by the mean intensity value averaged over the primary sensorimotor cortex region, which has been shown to improve FDG-PET based discrimination compared to other brain regions or the global metabolic mean [YLB+08]. Anatomically, the "Precentral_L, Precentral_R, Postcentral_L and Postcentral R" regions in the AAL brain template can be used as the primary

Table 4.7: Baseline information mean (std) of the patients. The protocol of ADNI diagnostic and image acquisition is described in appendix. CDR: Clinical Dementia Rating. CDT: Clock Drawing Test, five is the best score and zero is the worst (Note that the representation of CDT score is contrary to TUM CDT score, cf. Section 2.2). MMSE: Mini-Mental State Examination. ADAS: Alzheimer's Disease Assessment Scale. The higher the ADAS, the more severe of mental illness. More explanation is referred to http://www.adni-info.org/scientists/Pdfs/ADNI GeneralProceduresManual.pdf

	Subjects (female:male)	Age	CDR Total	CDT	MMSE	ADAS
$rac{\mathrm{MCI_{MCI}}}{\mathrm{MCI_{AD}}}$	45 (13:32) 32 (12:20)	76 (8.2) 75 (6.9)	1.97 (0.3) 1.95 (0.3)	\ /	27.4 (1.6) 26.7 (1.7)	15 (6.6) 19 (4.6)

sensorimotor cortex. The analysis is performed only on the AAL defined brain regions (gray matter voxels of MNI space). In total, there are 185,405 voxels in the AAL brain template. Among 185,405 voxels in each FDG-PET scan, there is a portion of voxels that contain discriminative information. Hence, the following method is proposed to discover the discriminative voxels.

4.2.2.3 Identification of Discriminative Voxels

Cox regression is suitable to be applied when some censored records occur in a dataset, which is exactly the scenario of the ADNI follow-up study (end of study after, for example, five visits). Therefore, we use this technique in a first step to select the discriminative voxels of FDG-PET scans.

4.2.2.4 Introduction to Cox Regression (Survival Analysis)

Cox regression is a semi-parametric survival analysis method. It makes no assumption about the probability distribution of the survival time, assuming only proportional hazards. Cox regression is also known as the Cox proportional hazards model [Cox72]. The rate at which the failure happens or the patient suffers from a disease is known as the hazard function. Let $x_1, x_2, ..., x_p$ be the values of p covariates $X_1, X_2, ..., X_p$. The hazard function is defined as:

$$h(t) = h_0(t) \exp\left(\sum_{i=1}^p \beta_i x_i\right), \tag{4.12}$$

where β_1 , β_2 , ..., β_p is the $1 \times p$ vector of regression parameters and $h_0(t)$ is the baseline hazard function at that time. The coefficient vectors of the covariates are estimated using a maximum likelihood (ML) estimate, which is obtained by maximizing a partial likelihood function. The hazard function focuses on failing, whereas the survivor function focuses on "surviving" given survival up to a certain time point. The hazard function h(t) and survivor function s(t) can be derived from each other. The general formula for their relation is:

$$s(t) = \exp\left(-\int_{0}^{t} h(u)du\right). \tag{4.13}$$

Since we apply Cox regression on each single voxel in this first step, we have only one covariate β_1 to determine, and x_1 is the voxel intensity. Kleinbaum [KK11] offers a more comprehensive introduction to survival analysis.

4.2.2.5 Cox Regression Applied to FDG-PET

Given the baseline PET scan, we run the Cox regression on each of the 185,405 voxels (AAL defined) independently on 77 studied samples, collecting the ones that show a negative correlation β_1 with hazardness at a p-value smaller than 0.01 (i.e., voxels that correlate positively with the survival time). The statistical significance level of 0.01 instead of 0.05 is chosen because we apply no multiple comparisons correction. The reason why we did not perform multiple comparisons correction is that too many voxels may be discarded after correction. The resulting voxels are first filtered out in this way, before they are combined in a classification model (see below).

4.2.2.6 Selection of Neighboring Voxels by Infinite Gaussian Mixture Model

A number of discriminative (informative) voxels are identified by the Cox regression analysis. Since an informative voxel's neighboring voxels also tend to be informative (in part due to the partial volume effect [RME98]), we need to eliminate such an effect to avoid overfitting. We therefore applied the infinite Gaussian mixture model to divide the voxels into clusters based on their x, y, and z coordinates (the resulting clusters are illustrated in Fig. 4.9).

The infinite Gaussian mixture model (IGMM) [Ras12] was proposed as an extension of the widely applied Gaussian Mixture Model (GMM) [Bis06]. In a GMM, the number of clusters (components) is assumed to be fixed a priori, which is, in fact, hard to do in practice. Alternatively, one can apply model selection technique to further determine the number of clusters as illustrated in Section 4.1.2.2. On the contrary, IGMM assumes the number of clusters is unknown (infinity, not to limit the number), which is determined by the data in the end. The IGMM can be briefly written as follows:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k), \tag{4.14}$$

where μ_k , Σ_k and π_k are the mean, covariance and mixing proportion respectively. In addition, $\Sigma_{k=1}^K \pi_k = 1$, $\pi_k \ge 0$ and $\theta = \{\mu_k, \Sigma_k, \pi_k\}$. \mathcal{N} denotes the D-dimensional Gaussian distribution:

$$\mathcal{N}(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right). \tag{4.15}$$

By allowing $K \to \infty$, IGMM extends the GMM in terms of the number of clusters. The inference is achieved by Markov chain Monte Carlo (MCMC), performing Gibbs sampling for a number of iterations. In the end, the voxel data neighbouring each other are likely to be grouped in one cluster.

After applying IGMM, the clustered voxels in one cluster are adjacent to each other. We subsequently chose 10% of the voxels in each cluster to represent the respective cluster, selecting voxels spread widely across the cluster. Finally, we collected the chosen 10% of voxels in every cluster as the discriminative voxels for the prediction model. In our experiments, 10% empirically turned out to be optimal. The reason may be that, too few voxels, such as 1%, may exclude some discriminative ones hence causing underfitting. On the other hand, too many voxels, such as 50% may still cause overfitting. Another reason is that the complementary information conveyed by other non-imaging data (see Section "Building the Classification Model") cannot be fully used, due to the curse of dimensionality, if too many voxel features are selected in building the classification model. The workflow is depicted in Fig. 4.10.

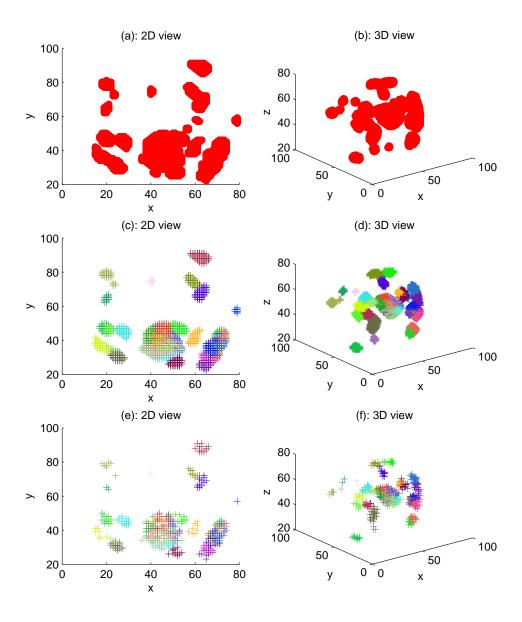


Figure 4.9: Illustration of significant voxels and the clustering results. (a): 3D view of the same voxels shown in 2D. (b): 3D view of significant voxels after Cox regression. (c): 2D view of clustered voxels (same color represents same cluster) after applying the infinite Gaussian mixture model. (d): 3D view of the same voxels shown in subplot (c). (e): 2D view of selected 10% voxels shown in subplot (c) and (d). (f): 3D view of the same voxels shown in subplot (e).

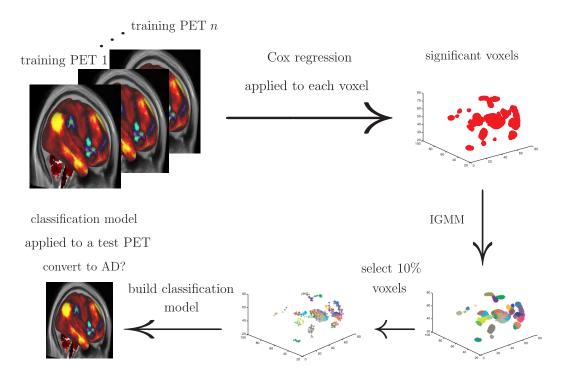


Figure 4.10: Workflow of proposed MCI conversion prediction. IGMM: infinite Gaussian mixture model.

4.2.2.7 Building the Classification Model

The identified imaging voxels were used as features to build a classification model. In this study, we employed support vector machines (SVM), which is a state-of-the-art classifier. LIBSVM [CL11] was used to build the SVM models with a linear kernel with grid search for parameter optimization. Grid search considers only the optimization of the penalty parameter in the linear SVM, selecting the value that yields the best classification result based on the training data. After the best value is found, it is applied to the test data. To validate the model, the whole dataset was split into two subsets, a training set used for model building and a test set used to test the performance of the model. A 5-fold cross-validation was applied to split the data, which was achieved by dividing it into five disjoint subsets, with four subsets as training and the remaining subset as test dataset. In addition to the imaging data (FDG-PET), we also investigated the non-imaging data (cf. Table 4.7) with the aim to consider information derived from more than just one source or modality.

The non-imaging data includes age, gender, the results of the Clock Drawing Test (CDT), the Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale (ADAS). Note that we did not use the Clinical Dementia Rating (CDR) in building the model, because CDR is a stronger indicator used for the diagnosis.

4.2.3 Experiments

The study samples show that the number of males is greater than the number of females in both MCI_{MCI} and MCI_{AD} (cf. Table 4.7). The two groups of MCI_{MCI} and MCI_{AD} indicate similar mean age and CDR at baseline. Other cognitive indicators, such as CDT, MMSE, ADAS, all suggest that the MCI_{MCI} group is at a better cognitive status than MCI_{AD} .

4.2.3.1 Brain Voxels Identified by Cox Regression Analysis

After applying a Cox regression analysis, a number of discriminative voxels served as essential features to train the subsequent model. Some identified voxels are displayed in Fig. 4.11. Table 4.8 reveals these voxels' associated regions defined in the AAL template. The "Precuneus_L, Precuneus_R, Parietal_Inf_L and Angular_R" account for particularly high percentages compared to other regions. It is known that the precuneus is involved in several essential cognitive tasks. For example, episodic memory, visual-spatial abilities, and motor activity coordination strategies. The parietal lobe includes symbolic functions in language and numbers and interpretation of spatial information. The remaining identified regions, such as angular, temporal and cingulum, are associated with some cognitive abilities as language, mathematics and memory, etc.

4.2.3.2 Prediction of Progression to AD Dementia

The baseline accuracy of random guessing is at 45/(45+32) = 58.4% (45 MCI_{MCI} and 32 MCI_{AD}) to predict that a patient does not progress to AD dementia within two years. Fig. 4.12 demonstrates that the FDG-PET scan alone achieves a classification accuracy of 70%, which is higher than the accuracy of any other information source. Classification accuracy reaches nearly 80% when all the sources of information are pooled together to build the predictive model, the reason being that the

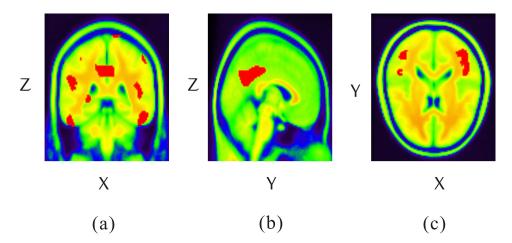


Figure 4.11: 2D view of the interesting regions at the 45th layer based on the whole dataset using Cox regression at a p-value of 0.01 (a): coronal view (b): sagittal view (c): transaxial view. x, y and z are the width (91), depth (109) and height (91), respectively. The red points represent the informative voxels (brain regions).

Table 4.8: The top-10 interesting regions and their proportions using the whole dataset. The proportion is calculated as the number of significant voxels in the region divided by the number of total significant voxels. The region name refers to the AAL template [TMLP⁺02].

Regions	Percentage	Regions	Percentage
1: Precuneus_L	13.9%	6: Temporal_Mid_L	5.79%
2: Precuneus_R	11.3%	7: Parietal_Sup_R	5.05%
3: Parietal_Inf_L	10.8%	8: Parietal_Inf_R	4.71%
4: Angular_R	10.3%	9: Parietal_Sup_L	3.84%
5: Angular_L	7.78%	10: Cingulum_Mid_R	3.57%

model benefits from gaining complementary discriminative information from diverse sources.

4.2.4 Discussion

Accurately predicting the course of MCI is an important clinical and academic aim, but censored data often limit the ability to derive meaningful results. Survival analysis is a suitable statistical tool for this kind of situation since it allows to analyze incomplete datasets. The current study corroborates findings of previous studies, showing that a metabolic deficit in the temporoparietal cortex, the precuneus and the

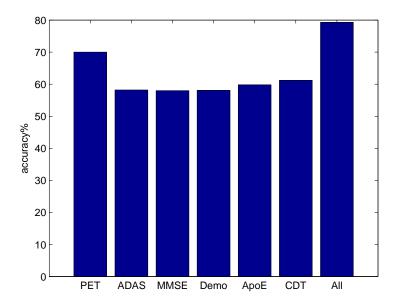


Figure 4.12: Comparison of classification accuracy of various information sources. All: using all data (PET, ADAS, MMSE, Demo, ApoE and CDT) as features in the SVM. ApoE: Apolipoprotein E.

limbic cortex offers valuable information in terms of dementia risk in MCI [MTH+08]. These brain areas are well known to play a role in episodic memory, visuospatial processing and executive function, which all are typically affected early in the course of AD [WWS12]. Our results also show that survival analysis is a viable statistical method to discriminate between progressive and stable MCI cases. The usual limitations of studies based on clinical cohorts recruited at specialized memory clinics apply to our study. These include the lack of generalizability of the results to the wider population due to highly selected MCI patients with a high a priori probability of suffering from AD; the lack of histopathological verification of the clinical diagnoses; the limited follow-up period; and the restricted sample size. We also demonstrate the benefits of a clustering algorithm (IGMM) to cluster the significant voxels derived from the Cox regression model. These voxels can be clustered into different groups with respect to their geometric similarity. By choosing a portion (i.e., 10%) of the voxels in each cluster, we avoid using too many voxels and thereby effectively reduce the risk of overfitting, while still maintaining informative voxels

for dementia prediction. We also corroborate the view that combining imaging data with neurocognitive and demographical information leads to improved classification accuracy. To combine various sources of data, one may also use multi-view stacking [LHS⁺11] (see also Chapter 3.3.2). However, the present research does not benefit from it because of the limited classification performance of non-imaging variables. In our case, collecting all data into a simple form for learning yields satisfactory results. The improvement in classification accuracy can be attributed to the complementary information provided by non-imaging variables, in other words, more complementary sources of information together improve the training model.

4.2.5 Conclusions

The present study proposed a survival analysis method to analyze neuroimaging data. It reveals the ability of FDG-PET to predict conversion from MCI to full-blown AD dementia within 24 months. By treating data from stable MCIs as censored records, we are able to use survival analysis to detect the brain regions that convey discriminative information. This approach shows brain regions that are typically associated with early clinical AD. Thus, it can be used for MCI to AD progression prediction. The use of IGMM divides the discriminative voxels into various clusters, and overfitting is eliminated by selecting a portion of voxels. Less discriminative power is conveyed by the neurocognitive and demographical data. They, however, still seem to provide complementary information, which altogether improves the accuracy of prediction model.

This page is intentionally left blank.

CHAPTER 5

Subgroup Discovery

5.1 Subgroup Discovery via Optimization

5.1.1 Motivation

Different subgroup discovery approaches differ in the way the search space is traversed and the way subgroups are evaluated (i.e., quality measures [LKF⁺04]). These aspects are not independent of each other, rather, they go hand in hand and directly determine the kind of solutions that are obtained. For example, some search strategies return subgroups with a certain degree of redundancy due to correlated attributes. Consider the example "IF weather = snow AND season = winter, THEN go skating". It is quite likely that information about "winter" is unnecessary once "snow" is identified. Since too many subgroups may incur the risk of overfitting as well as imply tedious work on the side of human users interpreting the results, we focus on the redundancy issue in this work and propose a novel search strategy to avoid redundancy based on quadratic programming.

The most simple and straightforward search method, exhaustive search, enumerates all possible combinations of conditions of attributes, however, it clearly becomes infeasible when the number of attributes is high. A second common way is beam search, where only a predetermined number of best partial solutions is evaluated at each stage. It is a heuristic technique, as it discards non-promising alternatives in order to keep the explored portion of the search space tractable. Applied to SD, it picks the top most promising subgroups for each level of the search and discards the

rest. On the other hand, standard pruning techniques cannot be used for subgroup discovery, due to the non-monotonicity of the quality function. For instance, subgroup $sd[x_1] = y$ may not be interesting, but its refinement may be interesting again. Further, $sd[x_1]$ and $sd[x_2]$ may not be interesting, but $sd[x_1, x_2]$ may still be. To deal with this, an optimistic estimate [Wro97] is usually employed, ensuring safe pruning of all refinements, if we know that all these refinements cannot pass the quality test. More recently, a tight optimistic estimate [GRW08] was proposed to speed up the search, which is also used for comparison in this work. The (tight) optimistic estimate is an (tight) upper bound of the quality of the promising subgroups. Having subgroups at hand, people usually gauge the quality score, significance, coverage, support, classification accuracy, rule complexity, and other measures [LKF+04].

The motivation of optimization based subgroup discovery is mainly two-fold: Runtime is a critical issue when we cope with high-dimensional datasets. Conventional methods, like beam search and optimistic estimate, may be computationally very expensive in such a case. Thus, we resort to an optimization technique to alleviate the hard combinatorial problem. Secondly, too many similar patterns would be too redundant and also laborious for end-users to comprehend. The proposed approach can identify correlated variables such that the resulting patterns can be less redundant but still predictive.

5.1.2 Subgroup Discovery via Quadratic Programming (SDVQP)

Conventional SD search methods, such as beam search and the optimistic estimate, evaluate subgroup rules *individually*, accompanied by two problems, i.e., the redundancy of subgroups and the nature of the exponential search space. The former is caused by treating each subgroup rule individually and generating rules that resemble each other and convey similar information. The latter is due to the inherent nature of the exponential search space. To reduce the redundancy and alleviate the combinatorial problem, we resort to an optimization technique (quadratic programming: QP) that iteratively discovers subgroup rules. At each iteration, QP selects a small number of useful attributes, which are subsequently analyzed in an exhaustive search. Then a number of rules will be applied and the samples covered by these rules are assigned new weights. The next iteration starts with a recalculated quality measure and the procedure is repeated until the stopping condition is met (cf.

Algorithm 5.1: Beam search

```
Data: F: set of feature values, d: maximum beam depth, k: beam width, R:
            subgroup rule, \rho: quality measure, D: qualified rule set.
 1 D \leftarrow \emptyset, cands \leftarrow \emptyset
 2 while depth \leq d \wedge improvement in beam do
        for i = 1 to k do
            R_{ij} \leftarrow \emptyset
 4
            for j = 1 to |F| do
 5
                R_{ij}: formed by adding a new feature value to old rule
 6
                cands \leftarrow cands \cup R_{ii}
 7
                if \rho(cands) \geq mimimum \ quality \ then
                   D \leftarrow D \cup R_{ij}
10
                end
            end
11
12
        end
        beam \leftarrow top k rules in cands
13
       depth + 1;
14
15 end
16 return D
```

Algorithm 5.2).

Generally, QP is a technique that leads to a desired solution given an objective and defined constraints, and thus especially suitable for yielding a globally (maybe locally) optimal solution in the presence of independent factors. The factors, in SD, can be attribute dependency and/or a quality measure. Attribute dependency naturally forms a quadratic term ($\mathbf{w}^T H \mathbf{w}$ in Eq. 5.2, "T" denotes matrix transpose), linking attributes in a pair-wise fashion⁵. The quality of individual attributes is viewed as a linear term ($Q^T \mathbf{w}$ in Eq. 5.2), representing the contribution of an attribute with respect to a given target. Therefore, the interaction between attributes as well as attribute contribution can be combined into one objective function that forms the core of the quadratic programming problem. The solution to the problem is a weight vector (\mathbf{w}) in which the non-zeros are further used for exhaustive search to actually find the SD rules. The identified attributes should have low correlation,

⁵ Partial correlation may be useful if we want to investigate the dependency between two variables after removing the effect of all others.

but at the same time high quality with respect to the target.

To be more precise, the quadratic program of SDVQP is defined as follows:

$$\begin{array}{ll}
 \underset{\mathbf{w} \in \mathbb{R}^{n \times 1}}{\text{maximize}} & \mathcal{F} = -\frac{1}{2} \mathbf{w}^T H \mathbf{w} + Q^T \mathbf{w}. \\
 \text{subject to} & \sum_{i=1}^n \mathbf{w}_i \le t \\
 & \mathbf{w}_i \ge 0,
\end{array} (5.1)$$

where $H \in \mathbb{R}^{n \times n}$ is the pair-wise mutual information matrix (symmetric, n being the number of attributes), since mutual information is non-negative, so $H_{ij} \geq 0$; t controls sparsity; $Q \in \mathbb{R}^{n \times 1}$ contains the quality score φ and is measured by the function WQM described in Algorithm 5.2. Specifically, if a certain value in an attribute has a support greater than 0.5 with respect to a target class, then it is regarded as an indicator for this target, thus its quality score is added to the total quality of this attribute $(Q_i = \sum_{j=1}^m \varphi(att_{ij}|supp(att_{ij} > 0.5)) \geq 0)$, where m is the number of distinct values in the ith attribute). In Eq. 5.1, the objective function \mathcal{F} is quadratic and the constraints are linear, thus it is standard quadratic program⁶ that can be solved by methods such as the interior point method. It is noteworthy that it is possible to weight Eq. 5.1 as $\mathcal{F} = -\frac{1}{2}(1-\alpha)\mathbf{w}^TH\mathbf{w} + \alpha Q^T\mathbf{w}$, $\alpha \in [0,1]$, allowing different contributions of the quadratic and the linear term. The proper value of α could be set via a cross-validation with respect to a quantity like classification accuracy. However, we set $\alpha = 0.5$ in this work.

Definition of Entropy:

$$H(X) = -\sum_{x \in X} p(x) \log p(x),$$
 (5.2)

where p(x) is probability distribution, log is logarithm with base 2 since information is transmitted in 0 and 1.

Mutual Information (MI) is a measure of the statistical dependence between two variables. Different from other measures like Pearson's ρ , MI can reflect also the non-linear relationship between variables. Let X and Y be two sets of discrete

⁶ Quadratic programming has been used for feature selection before by Rodrigue-Lujan et al. [RLHEC10].

Algorithm 5.2: Subgroup Discovery via Quadratic Programming (SDVQP) **Data**: SR: subgroup rules, d: maximum number of attributes in SR, ρ : quality measure, δ : quality threshold, \mathcal{S} : qualified subgroup rules, |att|: number of attributes, $|att_i|$: number of possible values (conditions) in ith attribute, N: number of samples. Initialization: $\mathbf{w} \leftarrow \arg\max_{\mathbf{w}} - \mathbf{w}^T H \mathbf{w} + Q^T \mathbf{w}, \ Z \leftarrow [1, ..., 1]^T \text{ of length } N, \text{ i.e.,}$ the initial weight for each sample is set to one, $iter \leftarrow 0$ while $\frac{|\mathcal{F}^{iter} - \mathcal{F}^{iter-1}|}{\mathcal{F}^{iter}} > tolerance \ and \ iter \leq maxIter \ \mathbf{do}$ $\mathbf{w} \leftarrow \arg\max_{\mathbf{w}} - \mathbf{w}^T H \mathbf{w} + Q^T \mathbf{w}$ select non-zero weights in \mathbf{w} as candidate attributes for subgroup discovery 4 using exhaustive search, if number of non-zero weights is greater than dthen select top-d $\mathcal{S} \leftarrow \mathcal{S} \cup SR \ (\varphi(SR) > \delta, \text{ Eq. } 3.6)$ 5 $M \leftarrow$ number of times each sample covered by SR6 for i = 1 to N do $Z_i = e^{-M_i}$ (new weights calculated for samples) 8 9 $Q \leftarrow \mathbf{WQM}(Z, data)$ 10 $iter \leftarrow iter + 1$ $\mathcal{F} \leftarrow \text{compute Eq. } 5.4$ **12** 13 end 14 WQM: weighted quality measure: **15 for** i = 1 *to* | *att*| **do** $q \leftarrow 0$ for j = 1 to $|att_i|$ do 17 if $supp(att_{ij}, Z) > 0.5$ then 18 $q \leftarrow q + \varphi(\operatorname{att}_{ij}, Z)$ 19 20 end $\mathbf{21}$

random variables, then MI is defined as:

 $Q_i \leftarrow q$

22 | (2 23 end

$$H(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x) p(y)},$$
 (5.3)

where p(x,y) is the joint probability distribution of X and Y, respectively, p(x)

and p(y) are the marginal probabilities. Eq. 5.3 can be rewritten as H(X;Y) = H(X) + H(Y) - H(X,Y), where $H(X) = -\sum_{x \in X} p(x) \log p(x)$ and H(X,Y) are the entropy of X and the joint entropy of X and Y respectively. If X and Y are independent $(X \perp \!\!\!\perp Y)$, then $H(X,Y) = H(X) + H(Y) \Rightarrow H(X;Y) = 0$. Once X and Y are totally dependent (e.g., identical), then $H(X,Y) = H(X) = H(Y) \Rightarrow H(X;Y) = H(X) \leq \log |\mathcal{X}|$, assuming that the random variable X takes on possible values \mathcal{X} , where equality is achieved only if X is a uniform distribution. After matrix multiplication, Eq. 5.1 can be equivalently expressed as:

$$\mathcal{F} = -\frac{1}{2}\mathbf{w}_{1}^{2}H_{11} - \dots - \frac{1}{2}\mathbf{w}_{i}^{2}H_{ii} - \mathbf{w}_{1}\mathbf{w}_{2}H_{12} - \dots - \mathbf{w}_{i}\mathbf{w}_{j}H_{ij} + \dots + Q_{1}\mathbf{w}_{1} + \dots + Q_{i}\mathbf{w}_{i}, \ (i, j \in [1, \dots, n]).$$
(5.4)

 H_{ii} in Eq. 5.4, self information, does not reflect the variable interaction, thus it seems that one can safely discard the terms $\frac{1}{2}\mathbf{w}_i^2H_{ii}$ to obtain a simplified version:

$$\mathcal{F}^* = -\mathbf{w}_1 \mathbf{w}_2 H_{12} - \dots - \mathbf{w}_i \mathbf{w}_i H_{ij} + Q_1 \mathbf{w}_1 + \dots + Q_i \mathbf{w}_i, (i, j \in [1, \dots, n], i \neq j).$$
 (5.5)

However, if we maximize \mathcal{F}^* instead of \mathcal{F} , we will arrive at a solution that selects always a single variable exhibiting maximal quality Q. Thus we have the following proposition.

Proposition 1: To maximize function \mathcal{F}^* , a full weight will be assigned to a single variable and all others to zero. Without loss of generality, we let $\sum_{i=1}^{n} \mathbf{w}_i = 1$.

Proof. Let Q_p be the maximal quality among Q_i (i = 1, ..., n), i.e., $Q_p > Q_q \ge 0$, $p \ne q, p, q \in [1, ..., n]$.

$$\mathcal{F}^* = -\mathbf{w}_i \mathbf{w}_j H_{ij} + Q_i \mathbf{w}_i \qquad (i, j \in [1, ..., n], i \neq j)$$
 (5.6)

$$\leq Q_i \mathbf{w}_i \qquad (\mathbf{w}_i, \mathbf{w}_j, H_{ij} \geq 0) \tag{5.7}$$

$$=Q_p\mathbf{w}_p+\ldots+Q_q\mathbf{w}_q$$
 (p, q takes on specific value of index i) (5.8)

$$\leq Q_p \mathbf{w}_p + \ldots + Q_p \mathbf{w}_q \tag{5.9}$$

$$= Q_p(\mathbf{w}_p + \dots + \mathbf{w}_q) \qquad \left(\sum_{i=1}^n \mathbf{w}_i = 1\right) \qquad (5.10)$$

$$=Q_{p} \tag{5.11}$$

Thus, the maximal value of \mathcal{F}^* is upper bounded by Q_p , hence a full weight is certainly given to $\mathbf{w}_p = 1$ to reach the maximum.

Consequently, only a single variable is chosen for subgroup discovery, which ends up with merely a single subgroup description (rule description with one attribute). However, subgroup discovery also aims at finding complex SD rules to gain some interesting insights, therefore maximizing Eq. 5.5 is not a preferable approach. In contrast to Eq. 5.5, Eq. 5.4 is a reasonable formulation for our purpose, because introducing $\mathbf{w}_i^2 H_{ii}$ can avoid assigning the full weight to a single attribute and ensures that the variable interaction term $\mathbf{w}_i \mathbf{w}_j H_{ij}$ plays some role in the optimization.

To maximize Eq. 5.4, the optimization program needs to select variables that together contain high quality (i.e., high Q_i) and with the least level of dependence (redundancy). In particular, non-zero weights of w are the ones that contribute the most to the objective function \mathcal{F} , therefore their corresponding attributes are selected for SD using exhaustive search. Since we intentionally impose t in Eq. 5.1 to be a small value, most of the weights are set to zero. Thus, it is realistic to apply exhaustive search on only a couple of attributes. In the first iteration, all samples are untouched and hence they are equally weighted. After the first iteration, some samples are covered by certain subgroup rules, and their new weights are computed as $Z = e^{-M}$, where M is the number of times an example has been covered. In such a way, unused samples contribute more to the score in Eq. 3.6. Alternatively, a multiplicative decrease $\frac{1}{M+1}$ can also be considered, in order to have a smoother drop of example weights. Consequently, also the number of discovered subgroup rules would be more than using the exponential reweighting. However, we do not re-weight samples in beam search (cf. Algorithm 5.1) and the optimistic estimate, because they will very likely deliver single rules (rules with only one attribute condition) and yield too few complex rules, in particular when using the exponential reweighting. As the sample weights are recalculated after each iteration, the quality measure Q (cf. line 10 in Algorithm 5.2) is immediately updated. However, we do not update the mutual information matrix H, because it encodes the variable interaction that

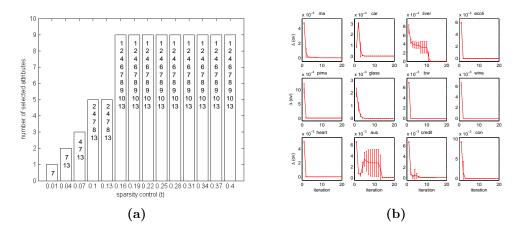


Figure 5.1: (a): Effect of sparsity control in the first iteration on the wine dataset. The numbers listed on the bar represent the selected attributes (i.e., $\mathbf{w}(att_i) \neq 0$), the order does not reflect their importance. Attribute 7 has the highest weight and hence is chosen first. (b): Objective function \mathcal{F} on 12 datasets with respect to the positive target class. Error bars indicate the standard deviation from 10-fold cross-validation.

indicates the inherent characteristics of the data. In addition, the algorithm needs to revise its coded information to guide the proper selection of subset attributes. Therefore, it should be kept unchanged.

Parameter Setting:

Parameter Setting I: sparsity control (t): The sparsity control parameter t was set to 0.1, Fig. 5.1(a) illustrates the impact of t with respect to non-zero weights ($\mathbf{w}_i \neq 0$). A larger t would return us more non-zero weights. Moreover, the choice of t is not particularly critical, because the most essential attributes always rank on top. Hence, even an inappropriately high t value still permits us to select the most influential attributes by picking the top ones. As shown in Fig. 5.1(a), a larger t always covers the selected attributes produced from a lower t, which suggests the consistency of variable selection yielded by the sparsity control parameter. To theoretically prove the consistency, we need to resort to the objective function (Eq. 5.1). Because the mutual information matrix H is not a positive definite matrix (cf. Proposition 2), Eq. 5.1 is not a convex function according to optimization theory, which indicates that a global maximizer might not be guaranteed, instead some local minimizer may be found. As a consequence, the theoretical proof of consistency of sparsity control is not assured. However, we empirically show that, in practice, the

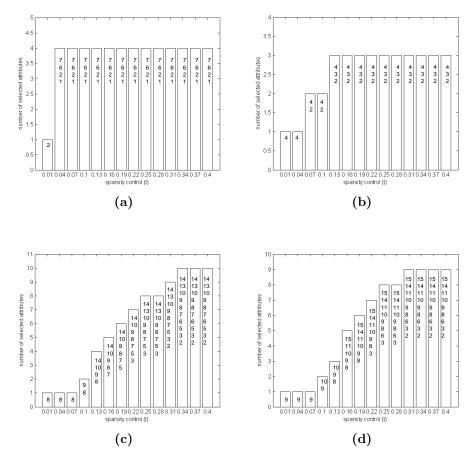


Figure 5.2: Effect of sparsity control. (a): mammography dataset. (b): breast cancer dataset. (c): australian statlog dataset. (d): credit dataset.

consistency is still well preserved on the randomly chosen four UCI datasets that are depicted in Fig. 5.2.

Parameter Setting II: termination condition (tolerance): When sample weights are newly weighted in each round, their corresponding counting statistics, such as support and quality, are generally decreasing, but the objective function value is not guaranteed to decline, since different attributes might be picked up at the next round and therefore an even higher value of \mathcal{F} may emerge. As a consequence, the value of $\frac{|\mathcal{F}^{\text{iter}}-\mathcal{F}^{\text{iter}-1}|}{|\mathcal{F}^{\text{iter}}|}$ (denoted as Δ (ov)) may not reach the specified tolerance (10⁻⁴) after several iterations. However, it will, in the long run, drop to a steady state as demonstrated in Fig. 5.1(b). Thus the algorithm terminates after the

maximum number of iterations maxIter. From Fig. 5.1(b), we observe that Δ (ov) reaches a stable state after 20 iterations. Thus, we may use 20 as a rule of thumb to set maxIter.

Proposition 2: Mutual information matrix is not positive definite.

Proof. Linear algebra states that a symmetric matrix H is positive definite if $x^T H x > 0, \ \forall x \neq 0.$ The mutual information matrix $H \in \mathbb{R}^{n \times n}$ is symmetric and its entry is denoted as H_{ij} . Let x be a $n \times 1$ column vector, $x \in \mathbb{R}^{n \times 1}$. Then, the matrix form of $x^T H x$ can be written as:

$$x^{T}Hx = \begin{bmatrix} x_{1} & x_{2} & \dots & x_{n} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1n} \\ H_{21} & H_{22} & \dots & H_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & H_{n2} & \dots & H_{nn} \end{bmatrix} \begin{bmatrix} x_{1} \\ x_{2} \\ \vdots \\ x_{n} \end{bmatrix}$$

$$= \begin{bmatrix} x_{1} & x_{2} & \dots & x_{n} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} H_{1i}x_{i} \\ \sum_{i=1}^{n} H_{2i}x_{i} \\ \vdots \\ \sum_{i=1}^{n} H_{ni}x_{i} \end{bmatrix}$$

$$(5.12)$$

$$= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n H_{1i} x_i \\ \sum_{i=1}^n H_{2i} x_i \\ \dots \\ \sum_{i=1}^n H_{ni} x_i \end{bmatrix}$$
(5.13)

$$= H_{11}x_1^2 + H_{12}x_1x_2 + \ldots + H_{1n}x_1x_n \tag{5.14}$$

$$+ H_{21}x_2x_1 + H_{22}x_2^2 + \ldots + H_{2n}x_2x_n + \ldots$$
 (5.15)

$$+ H_{n1}x_nx_1 + H_{n2}x_nx_2 + \ldots + H_{nn}x_n^2$$

$$= \sum_{i,j=1}^{n} H_{ij} x_i x_j. (5.16)$$

If n = 4, then $x^T H x$ can be expanded as $H_{11}x_1^2 + H_{22}x_2^2 + H_{33}x_3^2 + H_{44}x_4^2 +$ $2H_{12}x_1x_2 + 2H_{13}x_1x_3 + 2H_{14}x_1x_4 + 2H_{23}x_2x_3 + 2H_{24}x_2x_4 + 2H_{34}x_3x_4, \text{ since } H_{ij} = 0$ H_{ji} for symmetric matrix. It is sufficient to show a counter example to com-

plete the proof. For example, if a data set is $\begin{bmatrix} 3 & 4 & 2 & 5 \\ 1 & 4 & 4 & 4 \\ 5 & 4 & 4 & 3 \\ 1 & 3 & 2 & 2 \end{bmatrix}$ and the correspond-

ing mutual information matrix
$$H$$
 is
$$\begin{bmatrix} 1.5 & 0.3 & 0.5 & 1.5 \\ 0.3 & 0.8 & 0.3 & 0.8 \\ 0.5 & 0.3 & 1.0 & 1.0 \\ 1.5 & 0.8 & 1.0 & 2.0 \end{bmatrix}$$
 and the column vector
$$x = \begin{bmatrix} 7.2 & 5.4 & 4.9 & -9.9 \end{bmatrix}^T, \text{ then } x^T H x = -0.19 < 0, \text{ which violates the necessary}$$

 $x = \begin{bmatrix} 7.2 & 5.4 & 4.9 & -9.9 \end{bmatrix}^T$, then $x^T H x = -0.19 < 0$, which violates the necessary condition for a matrix being positive definite. Therefore the mutual information matrix is not positive definite.

5.1.3 Experiments

The different methods are performed on a set of benchmark UCI datasets [FA10] as well as a real-world medical dataset on Alzheimer's disease.

5.1.3.1 Experimental Results on UCI Dataset

Since subgroup discovery can be evaluated by different measures [LKF⁺04], we focus on cover redundancy, runtime, predictive accuracy and rule complexity. To measure predictive accuracy as well, we conduct a 10-fold cross-validation. The rules were built on both classes. To obtain a prediction for a new instance, the quality score ρ was summed for all the subgroups covering this instance for both classes. The class label with the highest overall quality is assigned. The algorithm can also be applied to multi-class problems if we perform one-versus-all or pairwise classification for the reduction of multi-class to binary. In the experiments, parameters were set as follows: a = 1 in Eq. 3.4, $tolerance = 10^{-4}$, $\delta = 0.01$ (quality threshold), and d = 4 (maximum number of conditions allowed in a subgroup rule).

As for cover redundancy CR, SDVQP nearly always halts with a lower value than other methods, which suggests a better rule diversity and therefore its resulting rule set is less redundant.

Results for the classification accuracy in Table 5.5 suggest that SDVQP performs better than beam search and the tight optimistic estimate. The mean classification performance on 12 datasets reaches 82.5%, which is the highest. Rule complexity is indicated by "avg(RF)", which measures how many attributes form a rule on average. Beam search and TOE subgroup rules have higher complexity than SDVQP rules, cf. Table 5.5. The results summary of Table 5.5 is provided in Table 5.2, in which

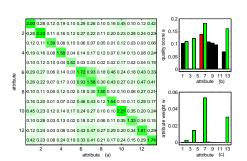
Table 5.1: Description of 12 UCI dataset [FA10]. †: samples with missing values were removed. ‡: multi-class datasets were converted to binary by merging several classes to one, i.e., the largest versus the rest. The continuous attributes were discretized by entropy-based discretization.

Dataset	# Samples	# Attributes	# Classes
mammography [ma]†	830	5	2
carţ	1728	6	4
liver disorder [liver]	345	6	2
ecoli‡	327	7	5
pima	768	8	2
glass‡	214	9	6
breast cancer wisconsin [bw]†	699	9	2
wine‡	178	13	3
heart statlog [heart]	270	13	2
australian statlog [aus]	690	14	2
credit approval [credit]†	653	15	2
congressional voting [con]†	232	16	2

the "RF" is not reported because it is hard to judge whether a short rule or a long rule is better.

In terms of runtime, SDVQP clearly outperforms the other approaches in the comparison, especially on higher-dimensional datasets, cf. Fig. 5.4. In beam search, as the beam becomes larger, the search space expands as well so that also more time is required. As for TOE, candidate generation is a time-consuming step at each level. Candidates for the next level are generated from qualified ones from the current level, which sometimes can cause memory problems.

Finally, Fig. 5.3 gives an example of how redundancy is handled by SDVQP. Fig. 5.3(c) shows that attributes 2, 4, 7 and 13 are selected, since they have non-zero weights and also exhibit a high quality score, as shown in Fig. 5.3(b). Interestingly, attribute 6, marked as red, has higher quality than 2 and 4, but is not chosen. The right-hand side table indicates that attribute 6 has high interaction with 2, 7 and 13, which are already picked, and therefore attribute 6 is ruled out. If it were chosen, some degree of redundancy would be introduced into the final subgroups. Eq. 5.4 makes sure that not only high quality contribution is considered, but also



2 4 7 13 1 Att 11 13 •6 1 MI 0.28 0.24 •0.93 0.42 Att 1 1 12 7	1	7		2		
MI 0.28 0.24 •0.93 0.42	_	•6				
MI 0.28 0.24 •0.93 0.42	0.42	-0	13	11	Att	1
2 Att 1 1 12 7		•0.93	0.24	0.28	MI	1
∠	7	12	1	1	Att	2
MI 0.28 0.19 0.47 0.41	0.41	0.47	0.19	0.28	MI	2
3 Att •6 7 9 •6	•6	9	7	•6	Att	2
MI •0.27 0.17 0.43 •0.33	0.33	0.43	0.17	•0.27	MI	3
4 Att 12 2 13 12	12	13	2	12	Att	4
MI 0.24 0.16 0.41 0.29	0.29	0.41	0.16	0.24	MI	4

Figure 5.3: Illustration of handling correlated attributes in SDVQP using the positive class of the wine dataset. (a): MI matrix. (b): attribute quality in the first iteration. (c): computed attribute weight (\mathbf{w}). The columns in the table are selected attributes (cf. Fig. 5.3(c)) and the rows are those attributes that share high MI with them, which are ranked in this order (cf. Fig. 5.3(a)).

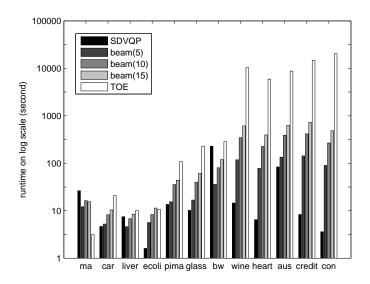


Figure 5.4: Runtime comparison on 12 datasets on log scale. Beam search with width 5, 10 and 15 was tested. The code was implemented in Matlab and runs on a machine with Intel(R) Dual Core(TM) i5 CPU @2.53 GHz, 4GB of memory.

low interactions among the variables. Thus, if two attributes share a high MI value, they would overall decrease the score of the objective function \mathcal{F} . However, clearly, the issues of redundancy and feature correlation are quite subtle and need to be investigated further.

Table 5.2: Results summary of Table 5.5 with the comparison of SDVQP versus compared methods (times of Win, Tie and Loss). Win: if proposed method SDVQP is better than compared method at the significance level (p-value < 0.05). Tie: if no significant difference is reported. Loss: SDVQP is significantly worse than the compared method. The p-value is calculated by the corrected paired t-test tailored for comparing learning algorithms [BF04]. In beam search, the beam width was set to 5, 10, and 15, respectively. b5: beam search method with beam width 5; results were obtained using all the returned subgroup rules. b*5: beam search results were obtained using the same amount (top ones) of rules as in SDVQP. For example, suppose |SR(SDVQP)| = 20, then select top-20 rules in beam*5 for comparison. So is the same as beam*10 and so on. acc: accuracy%; avg(RF): average number of attributes covered by a rule in a rule set for both classes. CR: cover redundancy [vLK11].

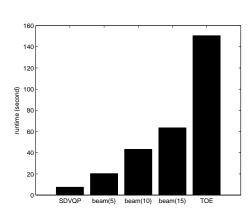
	1								TO D*
		b5	b*5	b10	b*10	b15	b*15	TOE	TOE*
	Win	4	2	4	2	4	2	5	2
acc	Tie	8	10	7	10	7	10	7	10
	Loss	0	0	1	0	1	0	0	0
	Win	11	11	11	11	11	11	11	11
CR	Tie	1	1	1	1	1	1	1	1
	Loss	0	0	0	0	0	0	0	0

Table 5.3: Description of Alzheimer's disease dataset. \overline{AD} : non-AD. Education is divided into four categories, "Realschule": 1, "Hauptschule": 2, "Gymnasium": 3, "Hochschule": 4. ApoE is of five types and their respective numbers are given in the third and fourth row. ApoE 23: ApoE E2/E3, etc.

class	sex	education	age	CDT	MMSE	ApoE
	(female, male)	(1, 2, 3, 4)				(23, 24, 33, 34, 44)
AD (23)	12, 11	9, 13, 1, 0	69	3.4	21	0, 1, 12, 7, 3
$\overline{\mathrm{AD}}$ (47)	26, 21	18, 25, 2, 2	67	2.1	26	5, 1, 16, 22, 3

5.1.3.2 Experimental Results on Alzheimer's Disease Dataset

The Alzheimer's disease dataset was provided by the psychiatry and nuclear medicine departments of Klinikum rechts der Isar of Technische Universität München. It comprises both imaging data and non-imaging data. The imaging data is in the form of F18-fluorodeoxyglucose positron emission tomography (PET) scans, which reflect the metabolic activity of the human brain and is an established biomarker of AD [Drz09]. We build on the data by Schmidt *et al.* [SHM⁺10] to evaluate our and



	acc	RF	CR
beam 5	85.1	2.95	-
beam* 5	86.3	2.73	0.47
beam 10	85.1	3.04	-
beam* 10	86.3	2.74	0.48
beam 15	85.1	3.12	-
beam* 15	86.3	2.74	0.48
TOE	85.1	3.39	-
TOE*	86.3	2.75	0.49
SDVQP	86.3	2.12	0.49

Figure 5.5: Results summary of Alzheimer's disease dataset. The figure depicts the runtime comparison of different methods. The table shows the same statistics as in Table 5.5. acc: accuracy%; avg(RF): average number of attributes covered by a rule in a rule set for both classes. CR: cover redundancy [vLK11]. beam 5: beam search method with beam width 5; results were obtained using all the returned subgroup rules. beam *5: beam search results were obtained using the same amount (top ones) of rules as in SDVQP.

competing methods. In this work, the PET scans were categorized into different clusters according to the intensity of the images. The optimal number of clusters was chosen according to the silhouette coefficient [KR90] based on the k-Medoids clustering method. We use the results yielded by k = 16 clusters. Since some of the clusters include too few cluster members, they are filtered out. The present study considers eight clusters, namely cluster 1, 4, 5, 6, 7, 9, 11, 14 and their respective numbers of PET images in these clusters are 4, 1, 4, 4, 0, 1, 6, 3 for AD and 3, 15, 14, 3, 2, 0, 3, 7 for non-AD.

On the other hand, the non-imaging data consist of demographic (e.g., gender, education, age) and clinical data. The clinical variables encompass two different psychometric tests, which are routinely used to assess a patient's overall cognitive performance. The first test is the clock drawing test, which requires the patient to draw a clock reading a specific time; it is scored between one (best) and six (worst). The second test is the MMSE, a battery of tests covering a range of cognitive domains with scores ranging from 30 (best) to 0 (worst). In addition, the Apolipoprotein E

Table 5.4: Selected subgroup descriptions (SD) using SDVQP. The score ρ is calculated as in Eq. 3.4.

```
SD1: IF MMSE \leq 25 THEN AD (\rho = 0.18)

SD2: IF MMSE \leq 25 AND ApoE = E3/E3 THEN AD (\rho = 0.08)

SD3: IF cluster = 11 THEN AD (\rho = 0.05)

SD4: IF CDT = 5 THEN AD (\rho = 0.02)

SD5: IF ApoE = E3/E3 THEN AD (\rho = 0.02)

SD6: IF MMSE > 25 THEN non-AD (\rho = 0.18)

SD7: IF CDT = 1 THEN non-AD (\rho = 0.08)

SD8: IF cluster = 4 AND MMSE > 25 THEN non-AD (\rho = 0.07)

SD9: IF MMSE > 25 AND ApoE = E3/E4 THEN non-AD (\rho = 0.07)

SD10: IF MMSE > 25 AND ApoE = E3/E3 THEN non-AD (\rho = 0.06)
```

genotype (ApoE) is also taken into account, since it is the strongest known genetic risk factor for sporadic, late-onset (AD).

In total, there are seven variables to investigate, namely image cluster, gender, education, age, CDT, MMSE and ApoE. We only consider the patients who have all this information, and therefore we end up with 70 patients' records, in which 23 are categorized as AD, and the remaining as non-AD, including individuals with only minor cognitive deficits (so-called mild cognitive impairment, MCI) or depression.⁷

The table in Fig. 5.5 suggests that the tested methods are comparable in terms of accuracy, rule complexity and cover redundancy. The main advantage of the proposed method is the runtime that is shown by the plot.

In Table 5.4, rule description 1 (SD 1) says that if the MMSE score is less than or equal to 25, then it is an AD patient, and this rule has the highest score value of 0.18. This rule is still qualified when it is complemented by ApoE E3/E3. The ApoE E3/E3 genotype is not associated with higher AD risk, but this combination passes the rule quality threshold due to the main contribution from feature MMSE ≤ 25 . SD3 describes that if the image is clustered into the eleventh cluster, then it is also an AD, which is consistent with the finding in the work of Schmidt *et al.* [SHM⁺10] (cf. Fig. 6 of that work). That is to say that the eleventh cluster comprises effectively a

⁷ Note that in contrast to a previous publication [SHM⁺10] the target variable is AD/non-AD, not the cluster membership in image clusters.

large portion of AD patients. SD6 reveals that a person is non-AD if the MMSE is greater than 25. Interestingly, SD9 tells that if MMSE is greater than 25 and and ApoE genotype is E3/E4, then it is a non-AD person. This particular genotype is associated with higher AD risk, but does not seem to provide relevant information in this case. However, a low quality threshold and a major contribution from another feature condition yields a controversial rule. For example, the rule condition MMSE is very dominant, which results in the controversial rule. The resulting rule set suggests that there are many rules of the same gist and thus many of them are redundant. We also see that the level of redundancy of the resulting rule set from the proposed method is not lower than the one of the competing methods. The reason may be that the limited number of samples is not representative enough, as opposed to the UCI datasets that comprise more samples. It would be interesting to see whether the interpretation of the subgroups would benefit from an increased sample size by data imputation. The pointed out facts remind us that the resulting rule set needs to be carefully examined by domain experts and the level of redundancy needs to be further coped with.

5.1.4 Conclusions

The work presented a subgroup discovery approach based on quadratic programming, aiming at reduced redundancy and improved computational efficiency. Instead of evaluating the subgroups individually, we utilize the mutual information matrix to explore the interaction between attributes. As a result, the degree of redundancy is reduced, which in turn avoids overfitting and thus makes classification more reliable, if used also in a predictive setting. Last, but not least, the proposed method runs much faster than other methods compared, which is a crucial factor when applied to high-dimensional data. As such, it offers an interesting alternative to beam search and the optimistic estimates, which have difficulty already on data of medium dimensionality. However, it should be kept in mind that the focus of this study was just on redundancy, computational efficiency, predictive power and rule complexity, whereas subgroups can be evaluated also along other dimensions.

Table 5.5: Results summary of UCI Datasets. Notations, such as b5 and b*5, are referred to Table 5.2. The best result is marked with bold face. Italics represent the salient difference between the compared methods and SDVQP. "-" denotes the empty value, since CR should be compared to rule sets with roughly the same size [vLK11]. "avg": averaged result over 12 datasets.

0	data	b5	b*5	b10	b*10	b15	b*15	TOE	TOE*	SDVQP
	ma	77.9	77.8	77.0	77.3	77.0	77.3	77.0	77.3	77.9
	car	88.3	79.5	86.2	79.5	86.2	79.	86.3	79.5	88.5
	liver	65.6	64.3	66.4	63.4	65.0	63.4	65.8	65.2	65.2
acc	ecoli	66.6	66.7	66.7	66.0	66.6	66.3	66.3	66.6	66.6
	pima	67.9	69.5	67.7	69.7	67.4	69.7	67.9	69.9	73.8
	glass	66.4	73.7	72.9	73.7	70.9	73.7	62.5	72.8	76.0
	bw	94.7	94.7	93.9	94.2	94.2	94.2	94.1	94.4	94.7
	wine	93.7	96.1	93.2	95.5	93.2	96.1	94.4	95.5	97.7
	heart	82.9	82.5	82.9	82.5	83.7	81.4	84.8	81.8	82.5
	aus	85.9	85.5	85.9	85.5	87.9	85.5	85.9	85.3	84.4
	credit	86.6	86.3	86.3	86.3	86.5	86.3	86.2	86.2	85.4
	con	94.4	97.0	93.1	97.0	92.7	97.0	90.5	97.0	97.0
	avg	80.8	81.1	81.0	81.0	81.0	80.9	80.1	81.0	82.5
	ma	2.72	2.33	2.72	2.35	2.72	2.35	2.73	2.36	1.84
	car	2.27	1.84	2.30	1.84	2.31	1.84	2.31	1.84	1.41
	liver	3.02	2.47	3.20	2.49	3.35	2.48	3.41	2.49	1.81
	ecoli	3.27	2.92	3.49	2.96	3.69	2.89	3.74	2.95	2.20
RF	pima	2.82	2.36	2.97	2.33	3.05	2.34	3.14	2.34	2.18
	glass	3.02	2.60	3.19	2.65	3.29	2.67	3.60	2.73	1.69
	bw	2.43	1.90	2.52	1.90	2.60	1.89	2.74	1.86	2.05
	wine	2.96	2.67	3.06	2.56	3.11	2.56	3.69	2.59	2.49
	heart	3.05	2.47	3.14	2.76	3.22	2.75	3.76	2.73	2.58
	aus	3.03	3.05	3.16	3.14	3.24	3.15	3.74	3.12	2.68
	credit	3.10	3.12	3.23	3.30	3.30	3.36	3.83	3.27	2.27
	con	3.17	3.05	3.26	3.28	3.31	3.32	3.92	3.32	2.35
	avg	2.91	2.56	3.02	2.63	3.10	2.63	3.38	2.63	2.13
	ma	-	0.34	-	0.32	-	0.32	-	0.33	0.46
	car	-	0.54	-	0.54	-	0.54	-	0.54	0.49
	liver	-	0.49	-	0.52	-	0.52	-	0.47	0.36
	ecoli	-	0.69	-	0.69	-	0.70	-	0.69	0.53
	pima	-	0.45	-	0.44	-	0.45	-	0.44	0.34
	glass	-	0.86	-	0.87	-	0.87	-	0.88	0.73
	bw	-	0.49	-	0.47	-	0.47	-	0.47	0.34
CR	wine	-	0.40	-	0.37	-	0.38	-	0.37	0.29
	heart		0.37	_	0.38		0.38	-	0.37	0.27
	aus	-	0.43	-	0.44	-	0.45	-	0.45	0.42
	credit	-	0.65	-	0.65	-	0.66	-	0.65	0.47
	con	-	0.34	-	0.34	-	0.34	-	0.34	0.19
	avg	-	0.50	-	0.50	-	0.51	-	0.51	0.40

5.2 Subgroup Discovery via Topic Modeling

5.2.1 Motivation

There are several important issues concerning subgroup discovery. First of all, the search strategy is an intensively studied topic, because the search space grows exponentially as the dimension increases. Thus, investigating all of the possible feature value combinations is simply infeasible for high dimensional data. To cope with it, beam search is used to explore only a tractable fraction of the search space. On the other hand, the optimistic estimate [GRW08, Wro97] is another alternative that discards the non-promising search branches and only concentrates on the top most promising subgroups at each level. The second essential aspect is the level of redundancy. During the process of subgroup rule mining, many similar rules can be found, although they all pass the selection criterion (e.g., quality measure). However, they may be some variants of the same scheme. Thus, discovering qualified but also redundancy reduced (diversity increased) subgroups is of great interest [LPDK15, vLK11]. Also, too many rules make it hard for users to interpret and validate the results. To address these issues of subgroup redundancy and interpretability, we approach the problem of subgroup discovery from a statistical perspective.

Motivated by the goal of rule interpretability, we conjecture it is easier to interpret rules once they are categorized, because categorization can reveal similarity/dissimilarity. In documents categorization, latent Dirichlet allocation (LDA) [BAJ03] is a generative topic modeling approach to identifying co-occurring words in documents. Each document can be characterized by a set of topics, and each topic is associated with a set of words. The popularity of LDA and its extension spreads across different application areas, such as document clustering, routine discovery, and so forth.

Subgroup discovery aims at finding conjunctions (co-occurrences) of feature values that together predict a target. On the other hand, LDA is meant to find co-occurring words in documents. Hence, both techniques uncover co-occurring patterns (words in LDA and feature values in SD). Thus, it is then feasible to lend the idea of LDA to SD to effectively find rules, without exhaustively searching the prohibitively large space of rules. Besides, a recent study [HS12] has shown that the use of

the Dirichlet process [Teh11], closely related to LDA, is efficient in finding frequent itemsets in binary transaction data. In addition, an Entity Topic Model (ETM) approach [KSHH12] was presented to devise topic models for documents with entity information by capturing the word co-occurrences. Inspired by this work, we present a *constrained latent Dirichlet allocation* (CLDA) approach to SD. Its main contributions are as follows:

- It offers another way of integrating LDA into SD to find interesting rules (a related method was proposed by Atzmüller and Mitzlaff [AM11]).
- A tailored CLDA is proposed to practically bring LDA and SD together.
- The resulting rules can be interpreted and categorized by various discovered topics, which is missing in existing SD algorithms.

5.2.2 Introduction to Topic Modeling

5.2.2.1 Fundamentals of Topic Modeling

In document classification, latent Dirichlet allocation (LDA) [BAJ03] is a generative topic modeling approach to identifying co-occurring words in documents. LDA is based on seminal work in latent semantic indexing (LSI [DDF⁺90]) and probabilistic LSI [Hof99]. In LDA, each document can be characterized by a set of topics, and each topic is associated with a set of words. The popularity of LDA and its extensions are applied in different application areas, such as document clustering, routine discovery [FGP11], and many others. To better explain LDA, it is necessary to introduce following statistical concepts as the Dirichlet distribution, conjugate priors and Markov chain Monte Carlo (MCMC).

Dirichlet distribution: It is a generalization of the Beta distribution, which is the case of n=2. It is a distribution overs distributions, i.e., a distribution over Multinomials. The Dirichlet distribution can be represented equivalently by Gamma random variables. The product of two Dirichlet distributions is still a Dirichlet distribution. Sampling from Dirichlet distribution amounts to sampling from independent Gamma distributions with common scale and shape parameters. It can be denoted as $x \sim \text{Dir}(\alpha_1, \dots, \alpha_n)$ or $x \sim \text{Dir}(\alpha)$ in the following form:

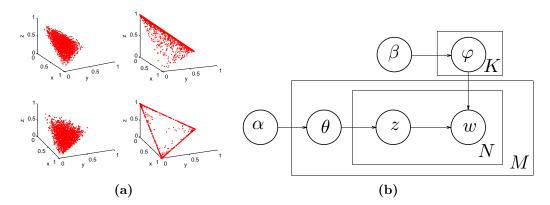


Figure 5.6: (a): Dirichlet distribution with $\alpha = [3\ 3\ 3], [0.1\ 1\ 1], [4\ 3\ 2]$ and $[0.05\ 0.05\ 0.05]$. (b): Graphical model of smoothed latent Dirichlet allocation [BAJ03]. θ is the per-document topic assignment and z is the per-word topic assignment. φ is the per-corpus topic distribution parameterized by β .

$$p(x;\alpha) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}, \quad \text{subject to } \sum_{i=1}^{n} x_i = 1, x_i \ge 0, \tag{5.17}$$

where ";" means x is parameterized by α . Γ is the Gamma distribution. Before proceeding to the conjugate prior, we introduce the multinomial distribution:

$$p(x; p, N) = \frac{N!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_i^{x_i}, \text{ subject to } \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} x_i = N,$$
 (5.18)

where $x_i \in \mathbb{N}^{+,0}$.

Conjugate prior: It is very useful in Bayesian statistics for inference. The conjugate prior facilitates the computation of posterior probability in a model, since the posterior form can be explicitly inferred. It can be shown that the Dirichlet distribution is the conjugate prior of the multinomial distribution, i.e., $p(x; p, N)_{mult} \cdot p(p; \alpha)_{Dir} = p(p; \beta)_{Dir}$, according to conjugate prior definition. The derivation is as follows:

Proof.

$$p(x; p, N)_{mult} \cdot p(p; \alpha)_{Dir} = \frac{N!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_i^{x_i} \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1}$$
(5.19)

$$\propto \prod_{i=1}^{n} p_i^{x_i} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \tag{5.20}$$

$$\propto \prod_{i=1}^{n} p_i^{x_i + \alpha_i - 1} \tag{5.21}$$

$$= p(p; x + \alpha)_{Dir} \quad \text{let } x + \alpha = \beta$$

$$= p(p; \beta)_{Dir}.$$
(5.22)

Thus, Dirichlet distribution is the conjugate prior of multinomial distribution.

It is worth mentioning that many well-known distributions such as Bernoulli, Poisson, Gaussian, Binomial, Multinomial and Dirichlet belong to the exponential family, which plays a key role in statistics. The exponential family has the general form:

$$P(x|\theta) = h(x) \exp\left\{\theta^T \psi(x) - A(\theta)\right\},\tag{5.23}$$

where $h(x) \ge 0$ is the base density, θ are natural parameters and $\psi(x)$ is the sufficient statistics vector. $A(\theta)$ represents the cumulant generating function or log partition function or log normalizer.

$$A(\theta) = \log \int \left(h(x) \exp\left\{ \theta^T \psi(x) \right\} dx \right). \tag{5.24}$$

 $A(\theta)$ plays an important role in exponential family distributions. In particular, it can be seen as the cumulant generating function of $\psi(x)$. The distribution in the exponential family is only a specific form of Eq. 5.23.

Markov chain Monte Carlo (MCMC): The aim of MCMC is to design a Markov chain whose stationary distribution is the target distribution. In Bayesian statistics, Gibbs sampling is a commonly used simple case of the MCMC sampling method. It is a random algorithm, and is an alternative to deterministic algorithms for statistical inference such as variational Bayesian or the expectation maximization (EM) algorithm. In LDA (cf. Fig. 5.6), we want to calculate the latent document-topic distribution φ , the topic-word distribution θ and the topic-word index assignment z. Due to the property of the conjugate prior, we can integrate out the multinomial parameters, simply sampling only z. This is called *collapsed* Gibbs sampling [GS04, Dar11]. The widely used Gibbs sampling is illustrated in Algorithm

Algorithm 5.3: Gibbs Sampling

```
1 Task: Sample joint distribution p(x_1, \dots, x_n).

2 Initialize x_1, \dots, x_n.

3 for i = 1 to MaxIteration do

4 | Sample x_1^{i+1} \sim p(x_1 | x_2^i, x_3^i, \dots, x_n^i).

5 | Sample x_2^{i+1} \sim p(x_2 | x_1^{i+1}, x_3^i, \dots, x_n^i).

6 | :

7 | Sample x_n^{i+1} \sim p(x_n | x_1^{i+1}, x_2^{i+1}, \dots, x_{n-1}^{i+1}).

8 end
```

5.3.

5.2.2.2 Latent Dirichlet Allocation

LDA is a generative model, describing how observed words in documents can be generated by hidden (latent) topics. The structure is shown in Fig. 5.6. In this model, there is a set of N words $\mathbf{w} = \{w_1, \dots, w_N\}$, M documents $D = \{d_1, \dots, d_M\}$, with each w_i belonging to some document d. In each document, there is a multinomial distribution over K topics, which are not directly observable (latent). Each document is associated with a parameter θ^d . A topic is also modeled as a multinomial distribution over words. Two hyperparameters α and β are introduced in this model. A low hyperparameter α encourages few topics per document and low β encourages few words per topic.

The LDA generative process can be explained as follows: Draw M multinomials θ from a Dirichlet prior α , one for each document; draw K multinomials φ from a Dirichlet prior β , one for each topic; draw a topic z_i from multinomial θ , i.e., $p(z_i|\alpha)$ and then draw a word w_i from multinomial φ , i.e., $p(w_i|z_i,\beta)$.

$$\theta \sim \text{Dirichlet}(\alpha)$$
 (5.25)

$$z_i | \theta^d \sim \text{Multinomial}(\theta^d)$$
 (5.26)

$$\varphi \sim \text{Dirichlet}(\beta)$$
 (5.27)

$$w_i|z_i,\varphi \sim \text{Multinomial}(\varphi_{z_i})$$
 (5.28)

The central task of inference in LDA is to determine the posterior distribution of latent topic variables θ and z conditioned on the words in the documents. Apart from the mean field variational methods [BAJ03], a collapsed Gibbs sampling approach was proposed to yield the solution [GS04]. The ratio of probability of w_i under topic j is:

$$p(w_i|z_i = j, z_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(\cdot)} + W\beta},$$
(5.29)

where $n_{-i,j}^{w_i}$ is the number of times a word assigned to topic j, excluding the current one, and $n_{i,j}^{(\cdot)}$ is the total number of times of words assigned to topic j, excluding the current one. W is the number of unique words. The probability of topic j in document d is:

$$p(z_i = j|z_{-i}) = \frac{n_{-i,j}^d + \alpha}{n_{-i}^d + K\alpha},$$
(5.30)

where $n_{-i,j}^d$ is the number of times a word from document d assigned to topic j, not including the current one, and n_{-i}^d is the total number of times a word in document d, excluding the current one. Thus the full conditional posterior distribution for z_i $(p(z_i = j | z_{-i}, \mathbf{w}))$ can be obtained as:

$$p(z_{i} = j | z_{-i}, \mathbf{w}) = \frac{p(z_{i} = j, w_{i} | z_{-i}, \mathbf{w}_{-i})}{p(w_{i})}$$

$$\propto p(z_{i} = j, w_{i} | z_{-i}, \mathbf{w}_{-i})$$

$$= \underbrace{p(w_{i} | z_{i} = j, z_{-i}, \mathbf{w}_{-i})}_{likelihood} \underbrace{p(z_{i} = j | z_{-i})}_{prior}$$

$$= \frac{n_{-i,j}^{w_{i}} + \beta}{n_{-i,j}^{0} + W\beta} \cdot \frac{n_{-i,j}^{d} + \alpha}{n_{-i}^{d} + K\alpha}.$$
(5.31)

5.2.3 Constrained Latent Dirichlet Allocation for Subgroup Discovery with Topic Rules

He and Shapiro [HS12] attempted to discover frequent itemsets in binary data using the Dirichlet process. Our proposed algorithm CLDA is not confined to binary data, and we are also able to identify topics indicating subgroup rules. We illustrate how to incorporate the idea of LDA into SD to find interesting subgroup rules.

5.2.3.1 Bringing SD and LDA Together via CLDA

In the proposed method, a "feature = value" expression functions as a word in the topic model, thus the number of total distinct feature values in the data amounts to the total number of words in LDA. We also assume that there are topics expressing some perspectives on the data. Thus, the subgroup rules can be immediately discovered after inferring the topics and their associated feature values. Feature values from the same feature may be grouped into the same topic, whereas the rule conditions in SD should be from different features. Therefore, we should effectively impose some constraints that encourage feature values from the identical feature to go into different topics. To this end, we propose a CLDA approach tailored for finding subgroup rules.

Recently, CLDA [ZLXJ11, AZC09] was suggested to allow the use of prior knowledge. The cannot-link and must-link constraints were realized by incorporating a term in Eq. 5.31. It can be shown [And10] that the conditional probability can be altered by multiplying a factor at the right-hand side of Eq. 5.31. Differing from their work, we suggest a different form of the constraint devised for SD. For example, we can intentionally multiply it with 0 if we knew a word belonging a topic j with probability 0. Similar to existing work [ZLXJ11, AZC09], we allow a soft constraint modifying Eq. 5.31 as:

$$p(z_{i} = j | z_{-i}, \mathbf{w}) = p(w_{i,j}^{k}) \cdot \frac{n_{-i,j}^{w_{i}} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{d} + \alpha}{n_{-i}^{d} + K\alpha},$$
(5.32)

where $p(w_{i,j}^k)$ denotes the prior probability of a feature value w_i from feature k belonging to topic j, and it is computed as:

$$p(w_{i,j}^{k}) = \begin{cases} 1 & \text{if } n_{-i,j}^{k} = 0, n_{i,j}^{k} = 0\\ 1 & \text{if } n_{-i,j}^{k} = 0, n_{i,j}^{k} \neq 0\\ \frac{1}{n_{-i,j}^{k} + n^{k}} & \text{if } n_{-i,j}^{k} \neq 0, n_{i,j}^{k} = 0\\ \frac{n_{i,j}^{k}}{n_{-i,j}^{k} + n_{i,j}^{k}} & \text{if } n_{-i,j}^{k} \neq 0, n_{i,j}^{k} \neq 0 \end{cases}$$

$$(5.33)$$

where $n_{i,j}^k$ is the number of times of feature value w_i from feature k belonging to topic j. $n_{-i,j}^k$ is the number of times this topic j already assigned to the feature k excluding the current w_i . n^k is the number of distinct feature values in feature

class feature 1 feature 2 T1 T2 T3 T4
$$+ A_1 \quad D_2 \quad A_1 \quad 0 \quad 2 \quad 0 \quad 20 \\ + C_1 \quad D_2 \quad B_1 \quad 0 \quad 0 \quad 4 \quad 10 \\ + B_1 \quad D_2 \quad C_1 \quad 0 \quad 0 \quad 5 \quad 30 \\ - C_1 \quad E_2 \\ \vdots \quad \vdots \quad p_{A_1,1}^1 = 1 \text{ (T1) } p_{A_1,3}^1 = \frac{1}{4+5+3} \text{ (T3)} \\ p_{A_1,2}^1 = 1 \text{ (T2) } p_{A_1,4}^1 = \frac{20}{10+30+20} \text{ (T4)}$$

Figure 5.7: Numerical demonstration of CLDA regarding the four cases in Eq. 5.33. Feature 1 constitutes three distinct values (A_1, B_1, C_1) , in which A_1 is used to show the calculation of $p(w_{i,j}^k)$ assuming four topics T1 to T4. Note the samples in the positive and the negative class are used separately.

Algorithm 5.4: CLDA for Subgroup Discovery with Topic Rules

Data: SR: subgroup rules, K: allowed maximal number of topics, training data $\mathcal{D}_{\text{train}}$, test data $\mathcal{D}_{\text{test}}$.

- 1 Data preparation for CLDA
- 2 for i = 1 to K do
- 3 Run CLDA on positive and negative samples from $\mathcal{D}_{\text{train}}$, respectively
- 4 Calculate the perplexity using Eq. 5.34 based on $\mathcal{D}_{\text{test}}$
- 5 end
- 6 Determine an appropriate number of topics \mathcal{K}_{best} based on the calculated perplexity for j=1 to \mathcal{K}_{best} do
- 7 Choose the corresponding features of co-occurring feature values produced from positive and negative samples respectively, as candidates to find SR on the training data $\mathcal{D}_{\text{train}}$
- 8 Collect the rules SR
- 9 end

k, which is used to act as a Laplace smoothing term. The essence of $p(w_{i,j}^k)$ is to encourage feature values from the same feature to fall into different topics by investigating the previous topic assignments. Fig. 5.7 demonstrates that calculation of $p_{A_1,j}^1$ is only involved with prior counting statistics of A_1, B_1 and C_1 , regardless of D_2 from the second feature. As for topic 1 (T1) and 2, no prior statistics of other feature values B_1 and C_1 is given, therefore the prior probability belonging to the topic is 1. As for topic 3, B_1 and C_1 are already assigned to it with 4 and 5 times.

Thus, its prior probability for this topic is only $\frac{1}{4+5+3}$, where 3 is the number of distinct feature values in feature 1, i.e., A_1, B_1 and C_1 . In terms of topic 4, the probability is proportional to its assignment 20 over the total assignments 60.

SD needs a class label (supervised) to find the rules, whereas LDA (unsupervised) does not request any class information. Thus, we divide the data into positives and negatives, constructing the CLDA based on data from each of the two classes, respectively. When built on either of the classes, CLDA produces the co-occurring feature values regarding the respective target class. It is equivalent to state that feature values tend to appear together in the positive or negative class, which is in line with the goal of finding rules pointing to a given target. In Algorithm 5.4, line 7 to line 8 are devoted to finding the actual SD rules with a fixed number of topics $\mathcal{K}_{\text{best}}$. For each topic, we have some feature values associated with integers indicating the number of assignments. The larger the number, the more frequently it appears in that topic, and of course zero means no occurrence. We then find the corresponding features of these feature values for exhaustive SD rule search using the quality function of Eq. 3.4. One can also only examine the combinations of these feature values for SD rules, but this may limit the number of discovered rules. In particular, we suppose that some features as a whole describe a certain topic, therefore we execute the search in a broader space.

Data Preparation for CLDA: Line 1 in Algorithm 5.4 prepares the data for running CLDA. If the data is numeric, then we first discretize them into nominal. The data may be denoted as integers, such as 1, 2, etc. Thus, two different features can have the same feature value of, for example, 1, but 1 in a feature is different from 1 in another feature. We, hence, intentionally denote each feature value uniquely to form a set of feature values (just as a vocabulary in documents). As a result, each sample is represented by some feature values drawn from the feature value set. In Fig. 5.7, for example, the set is A_1, B_1, C_1, D_2 ... for the positive class and C_1, E_2 ... for the negative class.

Choosing the Number of Topics: It is often hard to know the number of topics in advance. One common remedy known from language modelling is the use of per-word predictive perplexity (low values are suggested) as a measure of the likelihood of the model based on a held-out test set [BAJ03]. It is a measure of the generalization ability of the model on unseen data. Theoretically, one can choose the best number of topics according to the lowest perplexity. We applied the perplexity

suggested by Heinrich [Hei04], which can be briefly formulated as:

perplexity(
$$\mathbf{w}_{test}$$
) = exp $\left\{ -\frac{\sum_{d=1}^{\mathcal{D}_{test}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{\mathcal{D}_{test}} N_d} \right\}$, (5.34)

where $\sum_{d=1}^{\mathcal{D}_{\text{test}}} N_d$ is the total number of feature values in the test data. $p(\mathbf{w}_d)$ is calculated as:

$$p(\mathbf{w}_d) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} \varphi_{j,i} \cdot \theta_{d,j} \right)^{n_i^d}, \tag{5.35}$$

where n_i^d is the number of times feature value (word) i appears in test sample (document) d. In this study, $n_i^d=1$ because a feature can appear only once in a sample. $\varphi_{j,i}$ is calculated using Eq. 5.29 only from training data. $\theta_{d,j}=\frac{n_j^d+\alpha}{\sum_{j=1}^K n_j^d+K\cdot\alpha}$, where n_j^d is the number of times a feature value assigned to topic j calculated using Eq. 5.30. For details, see Eq. 93 of the original reference [Hei04].

5.2.3.2 Insights into Rules by Topics

The four evaluation measures allow a comparison among SD algorithms, while one merit of the proposed approach is that it offers the possibility of getting deeper understanding of rules by investigating the topics. To gain further insights into the rules in various topics, we suggest a measure of pair-wise distance between every two sets of rules in two topics. The distance is measured on the every pair of rules in the two topics. The rules in a topic are called a topic rule set. We define a rule distance rd for single rules r_1 and r_2 as:

$$rd(r_1, r_2) = \frac{\text{Hamming distance}(r_1, r_2)}{\max(|r_1|, |r_2|)}.$$
 (5.36)

The Hamming distance measures the bitwise difference between two rules, and the denominator ensures the measure is bounded by [0,1]. It measures the dissimilarity/distance, as opposed to JI, a measure of similarity. For example, $rd(\{A\}, \{B,C\}) = 2/2$ and $rd(\{A\}, \{A,B,C\}) = 2/3$. The calculated pair-wise distance can be used to show a dendrogram as in Fig. 5.10.

Table 5.6: Description of six UCI dataset [FA10]. †: samples with missing values were removed. Att.: attributes. ‡: multi-class datasets were converted to binary by merging several classes into one, i.e., the largest versus the rest. The continuous attributes were discretized by entropy-based discretization.

Dataset	#Samples	#Att.	#Classes
pima	768	8	2
glass‡	214	9	6
heart statlog [heart]	270	13	2
credit approval [credit]†	653	15	2
bank	4521	17	2
german credit [GC]	1000	20	2

5.2.4 Experiments

The algorithms are tested on six UCI datasets [FA10]. A 10-fold cross-validation was conducted to hold out some data for calculating perplexity. The tested number of topics (T) was from 5 to 100. The hyperparameters were set to $\beta = 0.1$ and $\alpha = 50/T$ (same as in previous work [FGP11, GS04, Wan09]), where T is the number of topics (i.e., the testing number i in Algorithm 5.4). These values of hyperparameters turned out to be suitable also in our tests. For the CLDA inference, we implemented a collapsed Gibbs sampling approach with 500 iterations. The threshold was $\delta = 0.01$ for the SD quality function. The rules were post-processed by the likelihood-ratio χ^2 test [LKF⁺04] at a significance level of 0.05. Four methods (cf. Fig. 5.9) are employed to compare with the proposed CLDA. These methods were already introduced. They represent a diverse set of methods regarding SD, e.g., optimistic estimate, redundancy reduction and diversity. In DSSD, default parameters were used, except minCoverage = 1 and maxDepth = 4. We only chose the "equal" rule descriptions to stay comparable with other methods. Certainly, there are many other SD algorithms that can be compared with, but the chosen ones are the most recent approaches.

5.2.4.1 Comparison with Baseline Random CLDA

We first empirically show that the proposed CLDA is feasible by comparing it with randomly chosen features, i.e., random CLDA. For each topic in CLDA, the al-

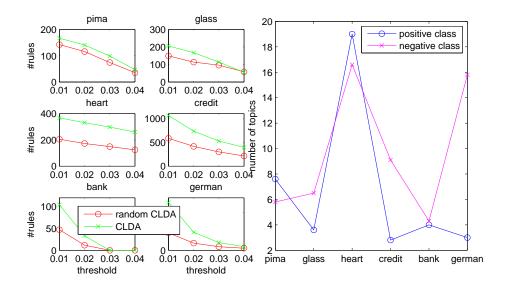


Figure 5.8: Left figure shows the relation between the number of rules and thresholds using *random* CLDA and CLDA methods. Right figure is the discovered actual number of topics on the six UCI datasets. The numbers are averaged over positive and negative classes from 10-fold cross-validation.

gorithm suggests some feature values co-occurring frequently, whose respective features are then used to identify the actual SD rules. Instead of using these identified co-occurring feature values, we *randomly* selected the same amount of feature values for rule search. Fig. 5.8 clearly shows that CLDA yields many more rules than random CLDA, which proves that CLDA can indeed find co-occurring feature values.

Fig. 5.8 illustrates that the number of topics is not influenced by the number of samples and dimensions. Regarding the positive class, heart reveals 19 topics, while german has only three topics despite it has the most samples and dimensions. The number of topics in the negative class varies slightly across these datasets.

5.2.4.2 Results on Six Datasets: Evaluation Measures

We focus on four measures (cf. Fig. 5.9) for our comparison, namely, cover redundancy, Jaccard index, accuracy and the number of rules. Concerning the Jaccard index (JI) measure, CLDA shows the lowest value on the pima, glass, heart and bank data. DSSD indicates a lower value on the remaining two datasets (credit and german), as it is particularly devised to discover *diverse* rules. As for redundancy,

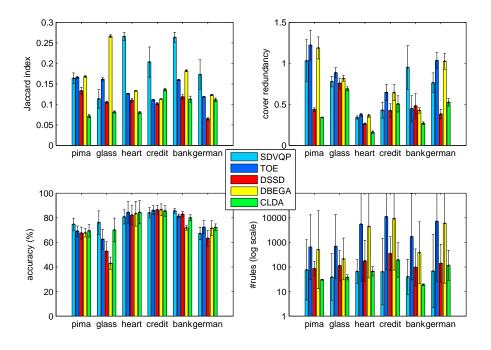


Figure 5.9: Evaluation measures of methods. The error bar represents the standard deviation from 10-fold cross-validation.

CLDA holds again the lowest value on pima, glass, heart and bank, being slightly worse (higher value) than DSSD on credit and german. The other three methods exhibit greater values than CLDA and DSSD overall. In terms of accuracy, all these methods show similar results, with SDVQP performing three times the best, on pima, glass and bank. The reason is that SDVQP integrates mutual information between feature and target class into the process of uncovering SD rules, therefore it has good predictive power. Regarding the number of discovered rules, TOE has the greatest number, since it finds all the qualified rules by shrinking the search space via an optimistic estimate [GRW08]. Next to TOE, DBEGA also returns more rules than the other three methods because it is a similar approach as TOE but focusing on generalization aware SD rules. SDVQP, DSSD and CLDA are not purely devised to find all SD rules, they are rather aiming for diverse rule sets. Thus, the size of the resulting rule set is smaller. In summary, CLDA returns rule sets within the same accuracy range as the other methods, but with comparatively low redundancy.

5.2.4.3 Results on Six Datasets: Insights into Rules by Topics

By design, CLDA also facilitates easier rule interpretation by categorizing rules into various topics. The dendrograms in Fig. 5.10 show that topics can be grouped by measuring their rules' similarity. If there are many rules, especially, we can interpret and examine these rules by looking at their topics. Choosing topics far apart in the dendrogram gives quite dissimilar rules, and choosing topics near each other gives similar rules. Hence, it is possible to interpret the SD rules via *uncovered hidden* topics. Take the glass dataset for example: Topic three (T3) and four (T4) are neighbors by sharing the second feature value marked by light green. In addition, CLDA gives a probability assignment to each of the feature values in every topic. This probability reveals how likely this feature value belongs to the topic.

5.2.5 Conclusions

This section presented a constrained latent Dirichlet allocation (CLDA) approach to discovering less redundant and more diverse subgroup rules. Instead of exhaustively searching the space of rules, we use a topic modeling method CLDA to identify co-occurring feature values. The feature values are associated with hidden topics, which are uncovered and used to find the actual SD rules. Consequently, the results revealed by the four evaluation measures indicate a better or similar performance compared to some standard methods.

In addition, the algorithm allows users not only pick the rules in terms of a rule quality measure, but also according to their associations to topics. The similarity of topics (hence rules) can be visualized by dendrograms using the suggested rule distance measure. Last, but not least, CLDA assigns a probability to each feature value in a discovered rule regarding the respective topic, which could aid users in gaining deeper insights into the data.

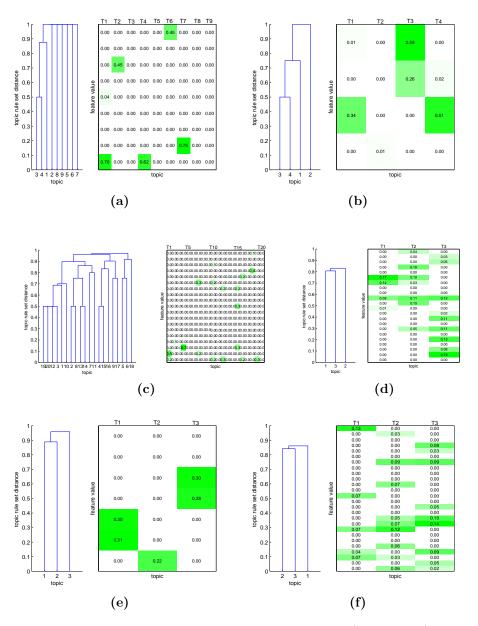


Figure 5.10: Dendrogram and calculated probability matrix (cf. Eq. 5.29) of feature values associated with yielded topics of the positive target class on six UCI datasets (cf. Table 5.6). Note that the sum of feature value probability in a topic is not one because the matrix shows only qualified SD rules.

This page is intentionally left blank.

CHAPTER 6

Multi-view Learning: Combining Imaging and Non-Imaging Data

6.1 Motivation

In Section 4, we proposed algorithms to mine imaging voxels. Apart from imaging techniques, clinical variables (e.g., from neuropsychological tests) are employed to assist in the assessment of cases [SHM+10]. As this medical domain naturally offers different perspectives or views on patients (e.g., demographic data, neuropsychological tests, imaging data, etc.), it lends itself to the application of methods for so-called multi-view learning [Re05]. In particular, we investigate the use of multi-view stacking, which combines the classifications (into the classes MCI or AD) originating from different views into overall classifications. In the proposed approach, we not only present the results of multi-view stacking along those lines, we also attempt to explain the performance of multi-view stacking in terms of correlations between feature groups and correlations between predictions. The proposed approach is demonstrated by the MCI against AD classification case. Finally, we shed some light on the way the different views are combined by the stacking procedure to come up with the overall classification.

6.2 Stacked Multi-View Learning

In this section, we introduce the proposed approach, which is slightly different from standard stacked generalization. Furthermore, we elaborate on the correlation measure using CCA that aims at providing a quantitative explanation of the results.

6.2.1 Stacking

Stacking [Wol92] was proposed to combine different classifiers to improve predictive accuracy. By learning how classifiers correlate with each other, the approach aims at outperforming each individual base classifier [FCS96]. Conventionally, base classifiers (originating from different base learners) are applied to a *single* dataset, and the predicted labels along with their true labels are concatenated and used as training (test) data at the meta level [Wol92, TW99]. In this study, stacking is applied in a different manner: we stack the predictions from different *views*, i.e., groups of features, and then perform meta level learning, assigning one base learner and classifier to each view. Each base classifier (from a corresponding view) then produces class probabilities (predictions), which are subsequently used to train the meta level model. Once a test sample is presented, each base classifier gives a prediction, and subsequently their predictions are combined by the meta model.

6.2.2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is applied to measure the correlation between views. It was proposed to measure the linear association between two sets of variables [Hot36]. Let $X=(x_1,x_2,\cdots,x_p),\ Y=(y_1,y_2,\cdots,y_q),\ U=\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_px_p=\alpha^TX$ and $V=\beta_1y_1+\beta_2y_2+\cdots+\beta_qy_q=\beta^TY$ (U and V are called canonical variates). $\Sigma=\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},\ \Sigma_{11}$ (or Σ_{22}) is the covariance matrix within set X (or Y), Σ_{12} is the covariance matrix between sets X and Y and $\Sigma_{12}=\Sigma_{21}^T$. CCA seeks to find α_p and β_q such that the following equation is maximized:

$$\rho(\alpha_p, \beta_q) = \alpha_p^T \Sigma_{12} \beta_q,$$
subject to $\alpha_p^T \Sigma_{11} \alpha_p = 1, \beta_q^T \Sigma_{22} \beta_q = 1.$

$$(6.1)$$

6.3 Experiments 107

Algorithm 6.1: Stacked Multi-View Learning.

```
Data: \mathcal{D}_i, i = 1, ..., n, n-view training data with true label Y.

1 Training: divide \mathcal{D}_i into J disjoint sets.

2 for i = 1 to n do

3 | for j = 1 to J do

4 | compute class probabilities z_{ij} of set j trained from remaining \mathcal{D}_i^{(J-j)} using base learner \mathcal{L}_i.

5 | end

6 | Denote z_i as entire predicted class probabilities of \mathcal{D}_i.

7 end

8 Let \mathcal{F} = \{z_1, \dots, z_i, Y\}, train meta level model (\mathcal{M}) based on \mathcal{F} using MLR.

9 Test: each \mathcal{L}_i gives a prediction (z_i^*) to a test sample \mathcal{D}_i^* trained from \mathcal{D}_i, let \mathcal{F}^* = \{z_1^*, \dots, z_i^*\} to be classified by \mathcal{M} and result in the final prediction.
```

As a result, the canonical correlation coefficient can be computed as the square root of the eigenvalues of matrix $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ for each canonical variates pair. The number of coefficients equals $min = \{p,q\}$ with a statistical significance value (p-value < 0.05), therefore $\mathbb{E}(\rho) = \sum_{i=1}^{min\{p,q\}} \rho_i / min\{p,q\}$ is taken as the overall correlation between two feature subsets. Pairwise CCA is performed among views and a mean value is calculated as the final correlation.

6.3 Experiments

The dementia dataset was provided by the psychiatry and nuclear medicine departments of Klinikum rechts der Isar of Technische Universität München. It covers 127 patients for which both a PET scan and clinical/demographic data are available. From these patients, 57 patients suffered from AD and 70 from MCI.

6.3.1 PET Imaging Data

Prior to their use, PET scans are transformed into feature vectors. The Automated Anatomical Labeling (AAL) brain Atlas [TMLP⁺02] was applied to obtain 116 predefined brain regions modeling the intensities of the interesting brain regions (e.g., Hippocampus) along with their spatial coordinates. This separation was done on a group of 20 cognitively healthy age-matched persons. We propose a density-based clustering method, in contrast to the model-based clustering method suggested in

Section 4.1.2.2. In particular, we applied DBSCAN [EKJX96] on the 116 previously identified regions and clustered them into 1894 finer groups⁸. Both intensity and coordinates (x, y, z) of each voxel were taken into account during the clustering. Then the mean intensity of each cluster of the MCI and AD PET scans was extracted. Thus, the PET is, at this point, represented as a feature vector consisting of 1894 intensity values. As 1894 features apparently require too much computational effort, we applied the F-score [CL06] to select the most informative ten features. However, before creating the final feature vector, PET images have to undergo two pre-processing steps: normalization and smoothing (kernel size [8 8 8] mm), which were achieved by SPM8⁹.

6.3.2 Non-Imaging Data

Non-imaging data consist of demographic (e.g., age and gender) and clinical data. Clinical variables cover neuropsychological tests that indicate the patients' social behavior, self-care capability and a person's daily ability concerning memory, language and orientation. The tests include: Consortium to Establish a Registry for Alzheimer's Disease (CERAD) neuropsychological assessment battery, Mini-Mental State Examination (MMSE) and the Clock Drawing Test (CDT). The CDT is added to the demographic data in this work to form a more informative view for multi-view stacking. The demographic data, CERAD, and the MMSE form natural views on the dementia data.

To examine the factors contributing to the success and failure of multi-view stacking, we additionally created 50,000 datasets with randomly generated views of the same dimensionality as of the natural views (e.g., $dim(\mathcal{V}_1) = dim(PET)$). From these 50,000 trials, we picked the two datasets with the lowest and the highest correlation among the views for further examination (cf. Table 6.1).

⁸ Parameter ϵ was analytically set according to the input data and the minimum number of points (minPts) = 6. The parameters were set so as to keep the clusters' size roughly balanced.

⁹ In our experiments, we found that SPM5 and SPM8 yield same results in terms of normalization and smoothing.

6.3 Experiments

6.3.3 Results

We ran three experiments to examine the prediction accuracy of stacking compared to a baseline learner. The baseline is a simple prediction based on the whole dataset without any division into views. K-nearest neighbors (KNN) is chosen as the base learner, as it is one of the simplest and most fundamental learning schemes¹⁰. KNN was used as first level predictor in the stacking approach, while multi-response linear regression (MLR) was applied as meta learner, since it was shown to be efficient for this purpose [TW99]. Table 6.1 gives the performance of the baseline approach (column "baseline"), predictions using only one view (columns "PET" to "Demo") and the accuracy for stacking ("Stacked"). Each experiment was repeated 50 times with a 10-fold cross validation. Class probabilities are used instead of class labels at the meta level. The K in KNN was set to nine and the probability for each class (AD and MCI) is calculated via majority voting of the first-level learner: $p(AD) = \frac{|\text{votes}_{AD}|}{K}$. Various values (K) have been tested, and values that are relatively large, e.g., nine, yield equally good results.

Table 6.1 shows that the natural views on dementia data present the best result for all stackings. Moreover, specific views outperform the baseline approach, but do not exceed the performance of stacking. Remarkably, all stacking approaches outperform the baseline approach. This shows that stacking on random views also increases performance. As for "Created Views 2", the result of stacking does not outperform \mathcal{V}_3 (82.7%), which may due to the too high base level correlation (0.95, cf. Table 6.2). Detailed explanation will follow in Section 6.4.

Although the stacked versions of Table 6.1 show that stacking is better than a baseline, the underlying mechanisms are not yet completely understood. Three parameters that may influence the performance of stacking are: the feature correlation between base views; the meta correlation, i.e., the correlation between the prediction of the separate views; and the variation of accuracy of separate views. Training data are used for these measures, since training data, in reality, are the only ones that we can gain knowledge from. Table 6.2 gives the values for the base correlation, meta correlation and the standard deviation of the accuracy of the four views. We claim that the prediction accuracy for stacking improves when the base level correlation is

¹⁰ Other learning schemes like decision trees or the SVM are also possible.

Table 6.1: Comparison of accuracy, mean \pm std%. Baseline: PET, MMSE, CERAD, CDT and Demographic data are pooled into one table for learning. \mathcal{V}_i : randomly created view of dementia data with dimension unchanged. In "Created Views 1", \mathcal{V}_2 was set to contain MMSE, since we wanted to observe if a strong individual view takes effect in multi-view stacking.

Natural Views	Baseline	PET	MMSE	CERAD	Demo	Stacked
KNN	76.3 ± 11.1	69.6 ± 12.4	78.1 ± 10.6	80.1±11.0	61.6 ± 13.4	83.2±10.2 •
Created Views 1	Baseline	\mathcal{V}_1	\mathcal{V}_2	\mathcal{V}_3	\mathcal{V}_4	Stacked
KNN	76.3 ± 11.1	72.9 ± 11.6	78.1 ± 10.6	79.6 ± 11.4	61.4 ± 12.9	81.2±11.0 •
Created Views 2	Baseline	\mathcal{V}_1	\mathcal{V}_2	\mathcal{V}_3	\mathcal{V}_4	Stacked
KNN	76.3 ± 11.1	75.6 ± 11.1	66.1 ± 11.8	$82.7{\pm}10.3$	72.0 ± 11.4	81.1±10.3 ●

Table 6.2: Description of three factors of stacked multi-view learning on dementia data. Base Corr.: base correlation, Meta Corr.: meta correlation, std of Accur: standard deviation of accuracy.

	Base Corr.	Meta Corr.	std of Accur.
Natural Views	0.63 ± 0.01	0.28 ± 0.04	13.9 ± 5.45
Created Views 1	$0.64 {\pm} 0.01$	$0.27{\pm}0.03$	13.5 ± 5.19
Created Views 2	$0.95{\pm}0.01$	$0.27{\pm}0.03$	12.2 ± 4.91

Table 6.3: Regressors of multiple linear regression on meta level data. p-value in brackets.

	PET	MMSE	CERAD	Demo
MCI	-0.241 (0.0001)	-0.303 (0.0031)	-0.269 (0.0157)	-0.083 (0.1791)
AD	0.339 (0.0006)	0.084 (0.0493)	0.197 (0.0108)	0.183 (0.2060)

relatively high, e.g., in the range of [0.6–0.9]. We can support this fact by examining the relation of the baseline correlation and the prediction improvement (for a more detailed analysis of these factors see Fig. 6.1 in Section 6.4). As the dementia dataset has a high baseline correlation, stacking is supposed to work on this data. As noted by other authors [FCS96], meta level correlation can be crucial in stacking, which is also supported by Fig. 6.1 that shows a likely ideal interval of [0.2–0.5]. Again, the given dementia dataset witnesses a meta correlation in this range. These findings, to some extent, should help explain the performance of stacked multi-view learning on our dementia dataset.

Another property of the views is, of course, their natural meaning. In the fol-

lowing, we will briefly analyze the contributions of the different views on the meta level classification. More specifically, we take a look at the coefficients of the MLR model, which is given as:

$$Y = \alpha \cdot \text{PET(MCI)} + \beta \cdot \text{PET(AD)} + \dots + \gamma \cdot \text{Demo(AD)} + \text{const.}$$
 (6.2)

As the MLR model encodes the two classes in two separate linear models, the task is to find eight regressors (weights) given the training label and meta level class probability. The resulting eight weights are given in Table 6.3. Following these results, MMSE and CERAD are strong at predicting MCI, since their weights are high (negative because MCI = 1 < AD = 2). By contrast, PET is strong at recognizing AD due to its highest value of 0.339. This table shows the interaction of these four views, which does not reveal their individual importance but their power in stacking using MLR as a meta learner.

6.4 Investigation of Stacking Performance

As it is hard to find standard datasets with naturally defined views, we conducted a study with synthetic views on 14 UCI datasets to empirically study the factors contributing to the performance of stacked multi-view learning.

6.4.1 UCI Data Subsets Generation

The UCI datasets are single view data, hence we randomly sample features into various subsets¹¹ (views) to create multiple views. The number of views produced is 4, which is the same as in the dementia data. 20 randomly sampled datasets are generated for every dataset, making sure that the generated views are approximately equally sized.

6.4.2 UCI Data Results

Three factors are analyzed, and the results shown in Table 6.4. The comparison with the dementia data is straightforward, because they are all of a 4-view scenario.

¹¹ UCI datasets were already split into two disjoint subsets for co-training purposes [LDZ09].

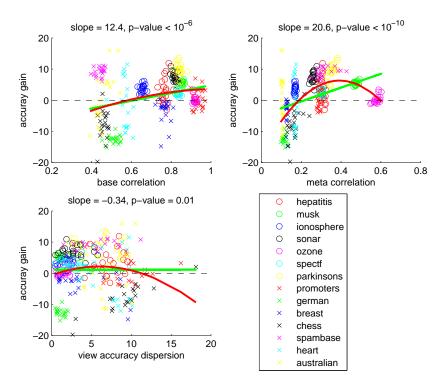


Figure 6.1: Regression analysis using all randomly sampled datasets. Each UCI dataset is re-sampled 20 times and each point is a dataset $(20 \times 14 = 280 \text{ in total})$. The green line is a linear regression, and the red line is a quadratic regression.

Table 6.4 reveals that the meta and base level correlation may be associated with the performance of stacking. For example, "german", "breast" and "chess" present low meta correlation and the stacked results are not satisfying, as opposed to "hepatitis", and "musk".

To measure the linear association between accuracy and the conjectured influential factors, we apply linear regression to determine the corresponding regressor (β) indicating their relation. Let "Y= accuracy gain = accuracy (stacking) – accuracy (baseline KNN)" be the dependent (target) variable, each of "base corr.", "meta corr." and "std of accur." be the independent variable X. The task is to find β given Y and X, assuming equation $Y=\beta\cdot X+$ const. The straight black line in Fig. 6.1 indicates the fitted linear curve with the slope and p-value shown on top. The gray quadratic curve is fitted using quadratic regression if we envision there is a non-linear relation.

Table 6.4: Accuracy comparison between baseline and stacked 4-view learning averaged over 20 randomly sampled sets (mean±std%). Only the mean is given for Base Corr. etc., as the standard deviation is marginal.

Datasets	KNN	Stacked 4-View	Base Corr.	Meta Corr.	std of Accur.
hepatitis [he]	56.7 ± 8.62	$62.6 {\pm} 3.92$	0.77	0.31	7.11
musk [mu]	71.0 ± 3.60	$85.7 {\pm} 0.62$	0.85	0.47	1.67
ionosphere [io]	85.1 ± 2.62	87.1 ± 1.39	0.66	0.17	1.93
sonar [so]	68.3 ± 5.27	$74.8{\pm}1.78$	0.82	0.26	2.93
ozone [oz]	93.3 ± 7.34	93.1 ± 1.30	0.92	0.58	1.49
spectf [sp]	73.8 ± 63.4	74.5 ± 0.94	0.83	0.30	2.62
parkinson [pa]	81.4 ± 6.79	83.1 ± 2.78	0.84	0.30	2.62
promoters [pr]	76.6 ± 6.09	77.6 ± 3.28	0.94	0.29	9.95
german [ge]	70.7 ± 2.19	50.9 ± 1.27	0.51	0.11	1.19
breast [br]	95.0 ± 1.47	90.2 ± 3.81	0.76	0.12	3.69
chess [ch]	91.5 ± 0.96	81.3 ± 4.56	0.46	0.21	9.26
spambase [sb]	91.0 ± 1.03	87.5 ± 1.47	0.44	0.30	5.52
heart [ht]	81.4 ± 3.51	63.1 ± 7.80	0.56	0.17	7.36
australian [au]	85.1 ± 2.30	69.3 ± 8.04	0.48	0.10	6.65

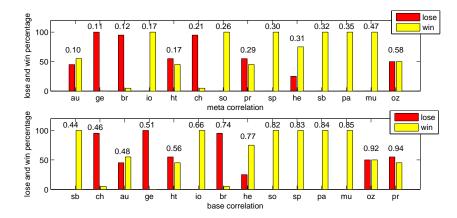


Figure 6.2: Percentage of wins (accuracy gain > 0) and losses (accuracy gain < 0) against meta and base level correlation. The correlation is shown on top of the bar. The number above each bar indicates the correlation noted in Table 6.4. For each dataset, the percentage of the yellow (win) and the red (lose) bar sums to one.

Fig. 6.1 demonstrates that meta correlation is by far the most important factor in terms of accuracy gain. A medium meta correlation would be suggested if the quadratic curve is regarded as more reasonable. The linear curve claims that accuracy grows as meta correlation increases. As for base correlation, the quadratic curve starts to drop at 0.9. View accuracy variation seems to be unimportant. Re-

markably, some datasets have high base and meta correlation but perform poorly. Therefore, we do not claim that the higher the correlation of base and meta level, the better the stacking performance, whereas a medium degree might be optimal. In Fig. 6.2, the majority of datasets with wins reveal relatively high base correlation [0.77–0.85] and medium meta correlation [0.30–0.47]. It can be easily understood that the meta level learner does not benefit if there is a very high or even perfect correlation. Certainly, there are still exceptional points falling into the optimal range, but behave poorly, which might be due to other reasons, such as the distribution difference between training and test samples, to name only one.

6.5 Discussion and Conclusions

The present study investigated the use of multi-view stacking for classifying dementia data, in particular for discriminating between Alzheimer's disease and mild cognitive impairment. While a simple KNN prediction including all features achieved a prediction accuracy of 76.3%, stacking on the natural views achieved an accuracy gain of 6.9%. Analyzing the meta level classifier showed that the predictions of the MMSE and CERAD views are important for MCI classification, while the PET view is crucial for AD. Further evaluations on 14 UCI datasets revealed that the performance can be largely attributed to the medium meta-level correlation of the views and the relatively high base-level correlation of the views. These insights were gained by transforming the UCI datasets into multi-view datasets within different ranges of base and meta correlations. Regression analysis showed that views with a high level of base correlation are likely to perform well in stacking. However, if the base correlation is low, stacking may nevertheless perform well. The same is true for the meta correlation: For medium values, the performance is likely to increase, but still this may happen for smaller and larger values. We can thus, to some extent, explain the good performance of stacking in our application domain. Based on these findings, researchers may therefore consider stacking for data from other medical application domains to improve their prediction accuracy whenever there are natural views in the indicated ranges of correlation.

This page is intentionally left blank.

CHAPTER 7

Conclusions and Outlook

7.1 Summary

In this thesis, we proposed algorithms to study structured patient data, with experiments both on real-world medical datasets and UCI datasets. Since imaging data (e.g., PET scans) are an important data source for clinical diagnosis, we investigate the use of PET from two perspectives. The first is dementia classification, which is achieved by clustering using a Gaussian mixture model with model selection. This approach extracts brain image features based on the resulting clusters. The number of clusters can be determined by the model selection, i.e., based on the Bayesian information criterion. The experimental results show that the method performs much better than compared methods especially on MCI against AD. The second application is the prediction of MCI progression, which is a follow-up study. We first apply survival analysis to identify some significant voxels. These voxels are subsequently clustered into a certain number of clusters by an infinite Gaussian mixture model. Further, a small portion of voxels from each cluster is chosen to build a classification model. The results reveal that PET has some power in predicting MCI progression. By adding other non-imaging variables to the model, we can obtain better results.

To extract some interesting rules from the data, we devised two new subgroup discovery algorithms. The optimization based approach pre-selects some potentially interesting features for further rule search. It leverages the feature's individual importance and the interaction between features. The resulting rule set is less 7.2 Outlook 117

redundant and more diverse as compared to other methods. The other approach is developed using a constraint latent Dirichlet allocation that has been proposed, which reveals various aspects (topics) in the data. Every topic is associated with some features that are then used for actual subgroup rule search. This method has one additional merit that the rules can be interpreted by the respective topic. A suggested dendrogram can illustrate the topic similarity in terms of contained rules.

A stacked multi-view learning approach is proposed to combine various information sources. The information contained in the sources can be better used when a two level stacking is performed. Stacking employs base level classifiers to learn the decision behaviors, then a meta level classifier makes better decisions based on the outputs of the base level. The experimental results suggest that combining various information sources can benefit the dementia classification performance. Experimental analysis explains the factors contributing to a good performance, i.e., that the correlation of views on the base and the meta level should be within certain ranges to facilitate successful stacked multi-view learning.

However, some limitations still exist in the work. The medical datasets do not contain many samples for a study, which complies with the common problem of small n large p. Because PET is an expensive neuroimaging technique, we can hardly obtain many PET scans to form a study. It is particularly more difficult to obtain follow-up PET scans for MCI progression studies. Other types of medical data are not abundant as well, such as biomarkers and neuropsychological test results. As a consequence, the developed subgroup discovery algorithms cannot be fully tested using the medical data.

7.2 Outlook

Regarding the research on imaging data, we may consider MRIs as another complementary source to PET. Currently, multi-modality is a popular topic in neuroimaging, because relying only on one information source may be error-prone and risky. However, this brings another research question of how to meaningfully combine various sources of information. We not only need methods like multiple kernel learning or stacking, but we also need to understand the underlying reasons of good performance and their applicable domains. Another essential aspect to be addressed is the missing records in medical data. It is nearly unavoidable to have missing values in

patient data, since patients may drop out of studies or some tests are simply not carried out. We encountered the problem of too few samples in the context of subgroup discovery for the medical data, which is a limitation of the experiments. Thus, it is desirable to further develop new algorithms that can cope with missing data, although some existing algorithms like expectation maximization can deal with it to some extent.

Although subgroup discovery has being actively studied in data mining for quite sometime, the issues of redundancy and rule search efficiency are yet to be well addressed. Hence, a good algorithm should return a comprehensible rule set as efficiently as possible. We may further extend the proposed constraint latent Dirichlet allocation to a non-parametric Bayesian approach in order to model rule discovery as a stochastic process. Such a process models data distribution, and the distribution can be nicely incorporated into a statistical model for learning. Therefore, there is no need to perform an exhaustive search on feature value combinations. In addition, the non-parametric Bayesian approach has the merit of avoiding model selection, which is instead replaced by model averaging. As a result, the number of topics does not need to be specified in advance. Therefore, the number of rules may be optimally determined by the approach based on the underlying data.

This page is intentionally left blank.

Appendices

Appendix

A.1 ADNI Data

ADNI Data Declaration:

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

122 A Appendix

ADNI Data Information Sharing Statement: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

ADNI Diagnostic and Image Acquisition Protocol: The ADNI recruitment and inclusion procedures are described in detail at www.adni-info.org. Briefly, at baseline, subjects in ADNI were between 55–90 years of age, had a modified Hachinski score e 4 and at least six years of education. Patients with MCI had MMSE scores between 24 and 30, a CDR score of 0.5; they had memory complaints, but no significant functional impairment, and objective memory deficits on the Wechsler Memory-Scale- Logical Memory II test. After the baseline visit, follow-up visits were conducted at six- or 12-month intervals up to a maximum of six years. FDG-PET were acquired within two weeks before or two weeks after the in-clinic assessments at Baseline and at the second annual visit, 24 months after Baseline.

A.2 Brain Regions defined in AAL

Table A.1: 116 defined brain regions in AAL brain template [TMLP⁺02].

	O	template [TMLP 02].
1: Precentral_L 2: Prec	central_R	3: Frontal_Sup_L
4: Frontal_Sup_R 5: Fron	$tal_Sup_Orb_L$	6: Frontal_Sup_Orb_R
7: Frontal_Mid_L 8: Fron	tal_Mid_R	9: Frontal_Mid_Orb_L
10: Frontal_Mid_Orb_R 11: Frontal_Mid_Orb_R	$ontal_Inf_Oper_L$	12: Frontal_Inf_Oper_R
13: Frontal_Inf_Tri_L 14: Fro	${ m ontal_Inf_Tri_R}$	15: Frontal_Inf_Orb_L
16: Frontal_Inf_Orb_R 17: Ro	$landic_Oper_L$	18: Rolandic_Oper_R
19: Supp_Motor_Area_L 20: Su	pp_Motor_Area_R	21: Olfactory_L
22: Olfactory_R 23: From	ontal_Sup_Medial_L	24: Frontal_Sup_Medial_R
25: Frontal_Med_Orb_L 26: Fro	$\operatorname{ontal_Med_Orb_R}$	27: Rectus_L
28: Rectus_R 29: Ins	$ m sula_L$	30: Insula_R
	ngulum_Ant_R	33: Cingulum_Mid_L
34: Cingulum_Mid_R 35: Cingulum_Mid_R	ngulum_Post_L	36: Cingulum_Post_R
37: Hippocampus_L 38: Hip	ppocampus_R	39: ParaHippocampal_L
	nygdala_L	42: Amygdala_R
43: Calcarine_L 44: Ca	lcarine_R	45: Cuneus_L
	ngual_L	48: Lingual_R
49: Occipital_Sup_L 50: Oc	cipital_Sup_R	51: Occipital_Mid_L
	$cipital_Inf_L$	54: Occipital_Inf_R
55: Fusiform_L 56: Fu	siform_R	57: Postcentral_L
<u>—</u>	rietal_Sup_L	60: Parietal_Sup_R
61: Parietal_Inf_L 62: Pa	$rietal_Inf_R$	63: SupraMarginal_L
64: SupraMarginal_R 65: An	gular_L	66: Angular_R
67: Precuneus_L 68: Pre	ecuneus_R	69: Paracentral_Lobule_L
70: Paracentral_Lobule_R 71: Ca	$udate_L$	72: Caudate_R
	$tamen_R$	75: Pallidum_L
76: Pallidum_R 77: Th	alamus_L	78: Thalamus_R
_	$schl_R$	81: Temporal_Sup_L
82:Temporal_Sup_R 83: Te	mporal_Pole_Sup_L	84: Temporal_Pole_Sup_R
85: Temporal_Mid_L 86: Temporal_Mid_L	$mporal_Mid_R$	87: Temporal_Pole_Mid_L
	mporal_Inf_L	90: Temporal_Inf_R
	rebelum_Crus1_R	93: Cerebelum_Crus2_L
	$rebelum_3_L$	96: Cerebelum_3_R
	$rebelum_4_5_R$	99: Cerebelum_6_L
	erebelum_7b_L	102: Cerebelum_7b_R
103: Cerebelum_8_L 104: C	erebelum_8_R	105: Cerebelum_9_L
	$ m erebelum_10_L$	108: Cerebelum_10_R
	ermis_3	111: Vermis_4_5
	ermis_7	114: Vermis_8
115: Vermis_9 116: V	ermis_10	

124 A Appendix

A.3 Derivation of Expectation Maximization Algorithm

The expectation maximization (EM) algorithm is a general method for deriving maximum likelihood parameter estimates from incomplete (i.e., partially unobserved) data. The essential trick of the EM algorithm is to maximize not the log-likelihood, but a lower bound on the log-likelihood, which is more tractable.

Let X be a random vector and parameterized by θ . We want to find θ such that $p(X|\theta)$ is a maximum, which is known as a maximum likelihood estimate (MLE) for θ . It is common to introduce the log-likelihood, since it is a strictly increasing function, i.e., $\mathcal{L}(\theta) = \ln p(X|\theta)$. Assume $\hat{\theta}_n$ is the n-th estimate of θ , then we want to maximize the difference,

$$\mathcal{L}(\theta) - \mathcal{L}(\hat{\theta}_n) = \ln p(X|\theta) - \ln p(X|\hat{\theta}_n). \tag{A.1}$$

In the presence of hidden random vector \mathbf{Z} with its given value \mathbf{z} , the total probability $p(X|\theta)$ can be written with respect to \mathbf{z} , $p(X|\theta) = \sum_{\mathbf{z}} p(X|\mathbf{z},\theta)p(\mathbf{z}|\theta)$, thus Eq. A.1 can be rewritten as:

$$\mathcal{L}(\theta) - \mathcal{L}(\hat{\theta}_{n}) = \ln\left(\sum_{\mathbf{z}} p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)\right) - \ln p(X|\hat{\theta}_{n})$$

$$= \ln\left(\sum_{\mathbf{z}} p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta) \frac{p(\mathbf{z}|X, \hat{\theta}_{n})}{p(\mathbf{z}|X, \hat{\theta}_{n})}\right) - \ln p(X|\hat{\theta}_{n})$$

$$= \ln\left(\sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_{n}) \frac{p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)}{p(\mathbf{z}|X, \hat{\theta}_{n})}\right) - \ln p(X|\hat{\theta}_{n})$$

$$= \ln\left(\sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_{n}) \frac{p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)}{p(\mathbf{z}|X, \hat{\theta}_{n})}\right) - \ln p(X|\hat{\theta}_{n})$$

$$\geq \sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_{n}) \ln\left(\frac{p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)}{p(\mathbf{z}|X, \hat{\theta}_{n})}\right) - \ln p(X|\hat{\theta}_{n})$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_{n}) \ln\left(\frac{p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)}{p(\mathbf{z}|X, \hat{\theta}_{n}) p(X|\hat{\theta}_{n})}\right)$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_{n}) \ln\left(\frac{p(X|\mathbf{z}, \theta) p(\mathbf{z}|\theta)}{p(\mathbf{z}|X, \hat{\theta}_{n}) p(X|\hat{\theta}_{n})}\right)$$

$$\stackrel{\text{def}}{=} \Delta(\theta|\hat{\theta}_{n}). \tag{A.2}$$

Further we define $l(\theta|\hat{\theta}_n) \stackrel{\text{def}}{=} \mathcal{L}(\hat{\theta}_n) + \Delta(\theta|\hat{\theta}_n)$. In order to achieve the greatest possible increase in the value of $\mathcal{L}(\theta)$, EM calls for selecting θ such that $l(\theta|\hat{\theta}_n)$ is maximized. The updated value is denoted as $\hat{\theta}_{n+1}$. Then we have,

$$\begin{split} \hat{\theta}_{n+1} &= \arg\max_{\theta} l(\theta|\hat{\theta}_n) \\ &= \arg\max_{\theta} \left(\mathcal{L}(\hat{\theta}_n) + \sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_n) \ln \frac{p(X|\mathbf{z}, \theta)p(\mathbf{z}|\theta)}{p(X|\hat{\theta}_n)p(\mathbf{z}|X, \hat{\theta}_n)} \right) \\ &\text{remove terms which depend only on } \hat{\theta}_n \\ &= \arg\max_{\theta} \left(\sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_n) \ln p(X|\mathbf{z}, \theta)p(\mathbf{z}|\theta) \right) \\ &= \arg\max_{\theta} \left(\sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_n) \ln \frac{p(X, \mathbf{z}, \theta)p(\mathbf{z}, \theta)}{p(\mathbf{z}, \theta)p(\theta)} \right) \\ &= \arg\max_{\theta} \left(\sum_{\mathbf{z}} p(\mathbf{z}|X, \hat{\theta}_n) \ln p(X, \mathbf{z}|\theta) \right) \\ &= \arg\max_{\theta} \left(\mathbb{E}_{\mathbf{z}|X, \hat{\theta}_n} [\ln p(X, \mathbf{z}|\theta)] \right). \end{split} \tag{A.3}$$

Then the expectation and maximization steps are easy to obtain:

- E-step: determine the conditional expectation $\mathbb{E}_{\mathbf{z}|X,\hat{\theta}_n}[\ln p(X,\mathbf{z}|\theta)]$.
- M-step: maximize this expression with respect to θ .

To infer the GMM solution, the E-step is to evaluate the posterior probability for the hidden cluster k given the current parameters $\theta = \{\mu, \Sigma, \pi\}$ at t iteration, i.e., $p_{ik}^t = \frac{\pi_k^t p(x_i | \mu_k^t, \Sigma_k^t)}{\sum_{k=1}^K \pi_k^t p(x_i | \mu_k^t, \Sigma_k^t)}$. Note, we denote x_i as a point in a random vector X. At M-step, this expression needs to be maximized by setting the first derivative to zero with respect to μ , Σ and π respectively. As a result, we obtain the final solutions as Eq. 4.4, Eq. 4.5 and Eq. 4.6.

List of Figures

2.1	Healthy and Alzheimer's disease brain	7
2.2	Brain regions and functionality	8
2.3	Example of clock drawing test. Image obtained from Master thesis [Sch07].	11
3.1	PET scanner, 3D brain, PET slices and NC, MCI, AD PET image	16
3.2	91 PET scan slices in transaxial view	17
3.3	Three views of PET scan and coordinate system	18
3.4	Pie chart illustration of subgroup discovery	25
3.5	Example of combinations of attribute-value conditions in finding sub-	
	group discovery rules	26
3.6	Demonstration of stacking.	34
4.1	BIC and AIC score displayed for the 10th (top) and 30th (bottom) bin	
	in 50 bin using pSMC normalization of the ADNI dataset	41
4.2	Generalization error (GE) by BIC and AIC using pSMC normalization	
	of the ADNI dataset	43
4.3	Number of features selected versus the classification accuracy using	
	pSMC normalization of the ADNI dataset on MCI against AD	46
4.4	Relation between the number of clusters and the number of bins on ADNI	
	and TUM datasets based on the normal control PET scans	50
4.5	ROC curve of compared method on ADNI dataset	53
4.6	ROC curve of compared method on TUM dataset	54
4.7	The informative regions (voxels) of MCI against AD using ADNI dataset	
	of the 45th layer	55
4.8	The informative regions (voxels) of MCI against AD using TUM dataset	
	of the 45th layer	56

List of Figures 127

4.9	Illustration of significant voxels and the clustering results	64
4.10	Workflow of proposed MCI conversion prediction. IGMM: infinite Gaus-	
	sian mixture model	65
4.11	2D view of the interesting regions at the 45th layer based on the whole	
	dataset using Cox regression at a p -value of 0.01	67
4.12	Comparison of classification accuracy of various information sources	68
5.1	Effect of sparsity control and Objective function $\mathcal F$ on 12 datasets with	
	respect to the positive target class	78
5.2	Effect of sparsity control	79
5.3	Illustration of handling correlated attributes in SDVQP using the positive	
	class of the wine dataset.	83
5.4	Runtime comparison on 12 datasets on log scale	83
5.5	Results summary of Alzheimer's disease dataset	85
5.6	Dirichlet distribution and graphical model of smoothed Latent Dirichlet	
	allocation	91
5.7	Numerical demonstration of CLDA regarding the four cases in Eq. 5.33.	96
5.8	Left figure shows the relation between the number of rules and thresholds	
	using random CLDA and CLDA methods	100
5.9	Evaluation measures of methods	101
5.10	Dendrogram and calculated probability matrix (cf. Eq. 5.29) of feature	
	values associated with yielded topics of the positive target class on six	
	UCI datasets (cf. Table 5.6)	103
6.1	Regression analysis using all randomly sampled datasets	112
6.2	Percentage of wins (accuracy gain > 0) and losses (accuracy gain < 0)	
	against meta and base level correlation	113

List of Tables

2.1	Various forms of dementia	9
2.2	Categories and subtests in CERAD	10
3.1	Difference modalities in radiology and nuclear medicine imaging	16
4.1	Participant characteristics and scanner type	38
4.2	Grade of evidence of the BIC difference [KR95]	41
4.3	Result summary of three different methods on ADNI dataset	51
4.4	Result summary of three different methods on TUM dataset. Results	
	from AIC are noted in brackets, following the BIC results	52
4.5	Informative regions (voxels) of MCI against AD using ADNI dataset	57
4.6	Informative regions (voxels) of MCI against AD using TUM dataset	57
4.7	Baseline information mean (std) of the patients	61
4.8	The top-10 interesting regions and their proportions using the whole dataset.	67
5.1	Description of 12 UCI dataset [FA10]	82
5.2	Results summary of Table 5.5 with the comparison of SDVQP versus	
	compared methods (times of Win, Tie and Loss)	84
5.3	Description of Alzheimer's disease dataset	84
5.4	Selected subgroup descriptions (SD) using SDVQP	86
5.5	Results summary of UCI Datasets	88
5.6	Description of six UCI dataset [FA10]	99
6.1	Comparison of accuracy, mean±std%	110
6.2	Description of three factors of stacked multi-view learning on dementia	
	data	110

List of Tables 129

6.3	Regressors of multiple linear regression on meta level data. p -value in	
	brackets	110
6.4	Accuracy comparison between baseline and stacked 4-view learning aver-	
	aged over 20 randomly sampled sets (mean \pm std%)	113
A 1	110 1 0 11 · · · · · · · · · · · · · · ·	100
A.1	116 defined brain regions in AAL brain template [TMLP ⁺ 02]	123

List of Symbols

- \mathbb{R} domain of real values
- \mathcal{L} log-likelihood
- \mathcal{N} D-dimensional Gaussian distribution
- Γ Gamma distribution
- Σ co-variance matrix
- μ mean value
- π mixing proportion in mixture model
- σ standard deviation (std)
- H mutual information matrix
- N number of data points (samples)
- X a vector, $X = \{x_1, x_2, ..., x_n\}$
- x a variable in a vector X
- Dir Dirichlet distribution

Index

Akaike information criterion (AIC), 21	Cox regression, 61
Alzheimer's disease (AD), 1 Alzheimer's disease assessment scale (ADAS), 10 Alzheimer's Disease Neuroimaging Initiative (ADNI), 13 Apolipoprotein E (ApoE), 8 association rule mining, 23 automated anatomical labeling (AAL), 18	dendrogram, 98 density-based spatial clustering of applications with noise (DB-SCAN), 108 diverse subgroup set discovery (DSSD), 31 expectation maximization (EM), 38 exponential family, 92
Bayesian information criterion (BIC), 21 canonical correlation analysis (CCA),	Gaussian mixture model (GMM), 37 Gaussian mixture model and model selection (GMM+MS), 37 Gibbs compling, 92
106	Gibbs sampling, 92 grand mean normalization, 18
cerebrospinal fluid (CSF), 11 clinical dementia rating (CDR), 10 clock drawing test (CDT), 10 computer-aided diagnosis (CAD), 11 conjugate prior, 91	infinite Gaussian mixture model (IGMM), 62 Jaccard index (JI), 29
consortium to establish a registry for Alzheimer's disease (CERAD),	KNN, 33
10 constrained latent Dirichlet allocation (CLDA), 95 cover redundancy (CR), 29	laten Dirichlet allocation (LDA), 93 library for support vector machines (LIBSVM), 51
//	magnetic resonance imaging (MRI), 15

132 Index

```
Markov chain Monte Carlo (MCMC),
                                            subgroup discovery via quadratic pro-
        90
                                                    gramming (SDVQP), 75
MCI_{AD}, 59
                                            support vector machines (SVM), 46
MCI_{MCI}, 59
                                            survival analysis, 59
mild cognitive impairment (MCI), 1
                                            tight optimistic estimate (TOE), 28
Mini-Mental State Examination (MMSE),
                                            topic modeling, 31
        10
minimum description length (MDL),
                                            UCI, 13
        31
                                            voxel-based morphometry (VBM), 19
multi-response linear regression (MLR),
        33
multi-view learning, 31
multiple kernel learning (MKL), 32
mutual information (MI), 74
normal control (NC), 1
perplexity, 97
positron emission tomography (PET),
        15
primary sensorimotor cortex normaliz-
        ation, 18
principal component analysis (PCA),
        20
quadratic programming (QP), 72
region of interest (ROI), 20
sequential minimal optimization (SMO),
        32
stacked generalization (stacking), 33
stacked multi-view learning, 106
statistical parametric mapping (SPM),
        17
subgroup discovery (SD), 21
```

This page is intentionally left blank.

- [AF00] J. Ashburner and K. J. Friston. Voxel-based morphometry: the methods. NeuroImage, 11:805–821, 2000.
- [AIS93] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.
- [Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [ALZ08] ALZ. The prevalence of dementia worldwide, Alzheimer's Disease International, http://www.alz.co.uk/adi/pdf/prevalence.pdf, 2008.
- [AM11] M. Atzmueller and F. Mitzlaff. Efficient descriptive community mining. In 24th International Conference of the Florida Artificial Intelligence Research Society. AAAI Press, 2011.
- [And10] D. Andrzejewski. Incorporating Domain Knowledge in Latent Topic Models. PhD thesis, University of Wisconsin-Madison, 2010.
- [AZC09] D. Andrzejewski, X. J. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32, 2009.
- [BA04] K. P. Burnham and D. R. Anderson. Multimodel inference understanding AIC and BIC in model selection. Sociological Methods & Research, 33:261–304, 2004.

[BAJ03] D. M. Blei, Ng. Andrew, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

- [BF04] R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *The 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*), pages 3–12. Springer, 2004.
- [BG09] M. Boley and H. Grosskreutz. Non-redundant subgroup discovery using a closure system. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 179–194. Springer, 2009.
- [BHAM12] N. I. Bohnen, K. Herlolz, Y. Anzai, and S. Minoshima. Effectiveness and safety of 18F-FDG PET in the evaluation of dementia: A review of the recent literature. *Journal of Nuclear Medicine*, 53(1):59–71, 2012.
- [Bis06] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.
- [BLJ04] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, 2004.
- [BM98] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.
- [BP01] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5:213–246, 2001.
- [BS04] S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 19–26, 2004.

[Bur98] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. data mining and knowledge discovery. 2(2):121–167, 1998.

- [BZ09] B. Bringmann and A. Zimmermann. One in a million: Picking the right patterns. *Knowledge and Information Systems*, 18:61–81, 2009.
- [CGT⁺11] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehericy, and M. O. Haber. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011.
- [CHC⁺12] C. Chu, A. L. Hsu, K. H. Chou, P. Bandettini, and C. P. Lin. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59–70, 2012.
- [CL06] Y. W. Chen and C. J. Lin. Combining syms with various feature selection strategies. *Feature Extraction, Foundations and Applications*, 2(3):315–324, 2006.
- [CL11] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.
- [CM09] M. Culp and G. Michailidis. A co-training algorithm for multi-view data with applications in data fusion. *Chemometrics*, 23:294–303, 2009.
- [CmJ09] M. Culp, G. michailidis, and K. Johnson. On multi-view learning with additive models. The Annals of Applied Statistics, 3:292–318, 2009.
- [CN89] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.
- [Cox72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

[CS99] W. W. Cohen and Y. Singer. A simple, fast, and effective rule learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 335–342. AAAI Press, 1999.

- [CUD08] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of 24th Con*ference on Uncertainty in Artificial Intelligence (UAI 2008), pages 88–96, 2008.
- [Dar11] W. M. Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. Technical report, Technical report University of Guelph, 2011.
- [DBU+09] P. Degroote, M. Briquet, K. Uytterhoeven, K. Lefever, T. Morel,
 C. Aerts, F. Carrier, M. Auvergne, and A. Baglinand E. Michel.
 Evidence for nonlinear resonant mode coupling in the beta cep star
 hd 180642 (v1449 aql) from corot space-based photometry. Astronomy & Astrophysics, 506:111-123, 2009.
- [DDF⁺90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- [DFJ⁺07] B. Dubois, H. H. Feldman, C. Jacova, S. T. Dekosky, P. Barberger-Gateau, and J. Cummings. Research criteria for the diagnosis of Alzheimer's disease: revising the nincds-adrda criteria. *Lancet Neurology*, 6(8):734–746, 2007.
- [DFJ⁺10] B. Dubois, H. H. Feldman, C. Jacova, J. L. Cummings, S. T. Dekosky, and P. Barberger-Gateau. Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurology*, 9:1118–1127, 2010.
- [DFU11] P. Dhillon, D. Foster, and L. Ungar. Multi-view learning of word embeddings via CCA. In *Proceedings of Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011)*, 2011.

[DL99] G. Z. Dong and J. Y. Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99*, 1999.

- [DL13] V. Dzyuba and M. V. Leeuwen. Interactive discovery of interesting subgroup sets. In *Proceedings of the 12th International Symposium on Intelligent Data Analysis*, 2013.
- [dLML⁺07] M. J. de Leon, L. Mosconi, J. Li, S. De Santi, Y. Yao, and W. H. Tsui. Longitudinal CSF isoprostane and MRI atrophy in the progression to AD. *Journal of Neurology*, 254(12):1666–1675, 2007.
- [DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39(1):1–38, 1977.
- [Drz09] A. Drzezga. Diagnosis of Alzheimer's disease with [18F] pet in mild and asymptomatic stages. *Behavioural Neurology*, 21(1):101–115, 2009.
- [Dz04] S. Dzeroski and B. zenko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54:255–273, 2004.
- [EKJX96] M. Ester, H. P. Kriegel, S. Jörg, and X. W. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), pages 226–231. AAAI Press, 1996.
- [EMC+13] A. E. Echeverria, M. McCurdy, R. Castillo, N. V. Ramos, W. Buckley, E. Castillo, P. Liu, J. Martinez, and T. Guerrero. Proton therapy radiation pneumonitis local dose-response in esophagus cancer patients. *Radiotherapy and Oncology*, 106:124–129, 2013.
- [FA10] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[FCS96] D. W. Fan, P. K. Chan, and S. J. Stolfo. A comparative evaluation of combiner and stacked generalization. In *Proceedings of AAAI-*96 workshop on Integrating Multiple Learned Models, pages 40–46, 1996.

- [FGP11] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1):1–27, 2011.
- [FP08] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2:916–954, 2008.
- [FRD11] R. Filipovych, S. M. Resnick, and C. Davatzikos. Multi-kernel classification for integration of clinical and imaging data: Application to prediction of cognitive decline in older adults. In *Proceedings of the Second International Conference on Machine Learning in Medical Imaging*, pages 26–34, 2011.
- [FWFN+10] A. M. Fjell, K. B. Walhovd, C. Fennema-Notestine, L. K. McEvoy, D. J. Hagler, and D. Holland. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. *Journal of Neuroscience*, 30:2088–2101, 2010.
- [GAH+13] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert. Rand forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65:167–175, 2013.
- [GGK+06] H. Gandhi, D. Green, J. Kounios, C. M. Clark, and R. Polikar. Stacked generalization for early diagnosis of Alzheimer's disease. In Proceedings of the 28th IEEE EMBS Annual International Conference, pages 5350-5353, 2006.
- [GLR⁺09] J. M. Górriz, A. Lassl, J. Ramírez, D. Salas-Gonzalez, C. G. Puntonet, and E. W. Lang. Automatic selection of ROIs infunctional imaging using Gaussian mixture models. *Neuroscience Letters*, 460(2):108–111, 2009.

[Gom14] J. J. Gomar. Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data. Alzheimer's & Dementia, pages 1–9, 2014.

- [GP11] H. Grosskreutz and D. Paurat. Fast and memory-efficient discovery of the top-k relevant subgroups in a reduced candidate space. In Proceedings of the 21st European Conference on Machine learning and Principles and Practice of Knowledge Discovery in Databases, pages 533–548. Springer, 2011.
- [GRW08] H. Grosskreutz, S. Rüping, and S. 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*, pages 440–456. Springer, 2008.
- [GS04] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings* of the National Academy of Sciences, 101(Suppl. 1):5228–5235, 2004.
- [GSR⁺11] J. M. Górriz, F. Segovia, J. Ramírez, A. Lassl, and D. Salas-Gonzalez. GMM based SPECT image classification for the diagnosis of Alzheimer's disease. *Applied Soft Computing*, 11:2313–2325, 2011.
- [HCGJ11] F. Herrera, C. J. Carmona, P. González, and M. J. Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, pages 495–525, 2011.
- [Hei04] G. Heinrich. Parameter estimation for text analysis. Technical report, 2004.
- [HFP+08] C. Habeck, N. L. Foster, R. Perneczky, A. Kurz, P. Alexopoulos, R. A. Koeppe, A. Drzezga, and Y. Stern. Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. *NeuroImage*, 40(4):1503-1515, 2008.
- [HK05] J. W. Han and M. Kamber. *Data Mining: Concepts and Techniques*, 2nd Edition. Morgan Kaufmann Publishers, 2005.
- [Hof99] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on

Research and Development in Information Retrieval, pages 50–57. ACM, 1999.

- [Hot36] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
- [HS12] R. F. He and J. Shapiro. Bayesian mixture models for frequent itemset discovery, arxiv, abs/1209.6001. 2012.
- [Kat93] R. Katzman. Revising the definition of Alzheimer's disease: a new lexicon. *Neurology*, 43:13–20, 1993.
- [KK11] D. G. Kleinbaum and M. Klein. Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health). Springer, 2011.
- [Klö96] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In Advances in Knowledge Discovery and Data Mining, pages 249–271, 1996.
- [KM02] W. Klösgen and M. May. Census data mining-an application. In *Mining Official Data*, pages 65–79, 2002.
- [Kou07] Z. Z. Kou. Stacked graphical models for efficient inference in markov random fields. In Proceedings of the 2007 SIAM International Conference on Data Mining, 2007.
- [KR90] L. Kaufman and P. J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Wiley, New York, 1990.
- [KR95] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
- [KSC+08] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131:681–689, 2008.

[KSHH12] H. S. Kim, Y. Z. Sun, J. Hockenmaier, and J. W. Han. ETM: Entity topic models for mining documents associated with entitiess. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, pages 349–358, 2012.

- [LAK14] R. Li, Z. Ahmadi, and S. Kramer. Constrained latent dirichlet allocation for subgroup discovery with topic rules. In *ECAI: 21st European Conference on Artificial Intelligence*, pages 519–524. IOS Press, 2014.
- [Lan10] S. M. Landau. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–238, 2010.
- [LBP13] F. Lemmerich, M. Becker, and F. Puppe. Difference-based estimates for generalization-aware subgroup discovery. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 288–303. Springer, 2013.
- [LDZ09] C. X. Ling, J. Du, and Z. H. Zhou. When does co-training work in real data? In Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pages 596–603. Springer, 2009.
- [LHS⁺11] R. Li, A. Hapfelmeier, J. Schmidt, R. Perneczky, A. Drzezga, A. Kurz, and S. Kramer. A case study of stacked multi-view learning in dementia research. In *Proceedings of the 13th Conference on Artificial Intelligence in Medicine: Lecture Notes in Computer Science*, pages 60–69. Springer, 2011.
- [LK12] V. Leeuwen and A. Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 2(25):208–242, 2012.
- [LKF⁺04] N. Lavrac, B. Kavsek, P. Flach, L. Todorovski, and S. Wrobel. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[LKTR12] L. Lu, N. A. Karakatsanis, J. Tang, and W. Chen A. Rahmin. 3.5D dynamic pet image reconstruction incorporating kinetics-based clusters. *Physics in Medicine and Biology*, 57:5035–5055, 2012.

- [LPDK15] R. Li, R. Perneczky, A. Drzezga, and S. Kramer. Efficient redundancy reduced subgroup discovery via quadratic programming. *Journal of Intelligent Information Systems*, 44(2):271–288, 2015.
- [LRG⁺] M. M. López, J. Ramírez, J. M. Górriz, I. Alvarez, D. Salas-Gonzalez, F. Segovia, and R. Chaves. SVM-based cad system for early detection of the alzheimer's disease using kernel PCA and LDA. Neuroscience Letters, 464:233–238.
- [LRG⁺09] M. López, J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia, and C. Puntonet. Automatic tool for Alzheimer's disease diagnosis using PCA and bayesian classification rules. *IET Electron*ics Letters, 45(8):389–391, 2009.
- [LZ14] E. Lammert and M. Zeeb. *Metabolism of Human Diseases*. Springer, 2014.
- [LZS12] M. H. Liu, D. Q. Zhang, and D. G. Shen. Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.
- [LZStADNI14] M. H. Liu, D. Q. Zhang, D. G. Shen, and the Alzheimer's Disease Neuroimaging Initiative. Identifying informative imaging biomarkers via tree structured sparse learning for AD diagnosis. NeuroInformatics, 12:856–867, 2014.
- [Met99] C. E. Metz. Evaluation of CAD methods. In Computer-Aided Diagnosis in Medical Imaging In: Doi, K.; MacMahon, H.; Giger, ML.; Hoffmann, KL., 251:543–554, 1999.
- [MF06] A. Martin and P. Frank. SD-Map a fast algorithm for exhaustive subgroup discovery. In *Proceeding of 10th European Conferences on Principles and Practice of Knowledge Discovery in Databases*, pages 6–17, 2006.

[MFNR⁺09] L. K. McEvoy, C. Fennema-Notestine, J. C. Roddey, D. Hagler Jr. D. J., Holland, D. S. Karow, C. J. Pung, J. B. Brewer, and A. M. Dale. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology*, 251:195–205, 2009.

- [MFS+00] V. Megalooikonomou, J. Ford, L. Shen, F. Makedon, and A. Saykin. Data mining in brain imaging. In Statistical Methods in Medical Research, pages 359–394, 2000.
- [MG13] S. Moens and B. Goethals. Randomly sampling maximal itemsets. In *IDEA: KDD Workshop on Interactive Data Exploration and Analysis*. ACM, 2013.
- [MMR88] J. C. Morris, R. C. Mohs, and H. Rogers. Cerad: clinical and neuro-psychological assessment of Alzheimer's disease. *Psychopharmacol Bull*, 4:641–652, 1988.
- [Mor93] J. C. Morris. The clinical dementia rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412–2414, 1993.
- [MS00] S. Morishita and J. Sese. Traversing itemset lattices with statistical metric pruning. In *Proceedings of the 19th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 226–236. ACM, 2000.
- [MTH+08] L. Mosconi, W. H. Tsui, K. Herholz, A. Pupi, A. Drzezga, G. Lucignani, E. M. Reiman, V. Holthoff, E. Kalbe, S. Sorbi, J. Diehl-Schmid, R. Perneczky, F. Clerici, R. Caselli, B. Beuthien-Baumann, and Minoschima S. de Leon M. J. Kurz, A. Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease, and other dementias. *Journal of Nuclear Medicine*, 49(3):390–398, 2008.
- [Mur12] K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

[MVT12] M. Mampaey, J. Vreeken, and N. Tatti. Summarizing data succinctly with the most informative itemsets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):16:1–16:42, 2012.

- [NLW09] P. K. Novak, N. Lavrac, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.
- [NSM+08] F. Nobili, D. Salmaso, S. Morbelli, N. Girtler, A. Piccardo, and A. Brugnolo. Principal component analysis of FDG PET in amnestic MCI. European Journal of Nuclear Medicine and Molecular Imaging, 35:2191–2202, 2008.
- [NST07] P. A. Naik, P. Shi, and C. L. Tsai. Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102:244–254, 2007.
- [PG14] J. Poirier and S. Gauthier. Alzheimer's Disease: The Complete Introduction (Your Health). Dundurn, 2014.
- [PKL⁺05] J. C. Price, W. E. Klunk, B. J. Lopresti, X. L. Lu, and J. A. Hoge. Kinetic modeling of amyloid binding in humans using PET imaging and pittsburgh compound-B. *Journal of Cerebral Blood Flow & Metabolism*, 25:1528–1547, 2005.
- [PZKM13] K. J. Preacher, G. J. Zhang, C. T. Kim, and G. Mels. Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48:28–56, 2013.
- [RÖ9] S. Rüping. Ranking interesting subgroups. In *Proceedings of the* 26th Annual International Conference on Machine Learning, pages 913–920. ACM, 2009.
- [Ras12] C. E. Rasmussen. The infinite Gaussian mixture model. In Advances in Neural Information Processing Systems, 12:554–560, 2012.

[RCG+09] J. Ramírez, R. Chaves, J. M. Górriz, M. López, I. A. Lvarez, D. Salas-Gonzalez, F. Segovia, and P. Padilla. Computer aided diagnosis of the alzheimer's disease combining spect-based feature selection and random forest classifiers. Nuclear Science Symposium Conference Record (NSS/MIC), 2009.

- [Re05] S. Rüping and T. Scheffer (editors). Proceedings of the ICML 2005 Workshop on Learning with Multiple Views. 2005.
- [RK08] U. Rückert and S. Kramer. Margin-based first-order rule learning.

 Machine Learning, 70:189–206, 2008.
- [RLHEC10] I. Rodrigue-Lujan, R. Huerta, C. Elkan, and C. S. Cruz. Quadratic programming feature selection. *Journal of Machine Learning Research*, 11:1491–1516, 2010.
- [RME98] O. G. Rousset, Y. L. Ma, and A. C. Evans. Correction for partial volume effects in PET: principle and validation. *Journal of Nuclear Medicine*, 39:904–911, 1998.
- [SAAZ12] M. S. Sharif, M. Abbod, A. Amira, and H. Zaidi. Artificial neural network-statistical approach for pet volume analysis and classification. Advances in Fuzzy Systems Special issue on Hybrid Biomedical Intelligent Systems, (5), 2012.
- [SBS⁺] F. Segovia, C. Bastin, E. Salmon, J. M. Górriz, and J. Ramírez. Combining pet images and neuropsychological test data for automatic diagnosis of alzheimer's disease. *PLoS ONE*, 464:233–238.
- [SCA⁺14] P. Y. Salaun, L. Campion, C. Ansquer, E. Frampas, and C. Mathieu. 18f-FDG PET predicts survival after pretargeted radioimmunotherapy in patients with progressive metastatic medullary thyroid carcinoma. *European Journal of Nuclear Medicine and Molecular Imaging*, 41:1501–1510, 2014.
- [Sch78] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[Sch05] M. Scholz. Knowledge-based sampling for subgroup discovery. In Local Pattern Detection, Lecture Notes in Artificial Intelligence, volume 3539, pages 171–189. Springer, 2005.

- [Sch07] J. Schmidt. Data mining in morbus Alzheimer data. Master's thesis, Technische Universität München, Germany, 2007.
- [SF05] J. Stoeckel and G. Fung. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 410–417, Washington, DC, USA, 2005. IEEE.
- [SFK10] P. Smialowski, D. Fishman, and S. Kramer. Pitfalls of supervised feature selection. *Bioinformatics*, 26(3):440–443, 2010.
- [SGR⁺12] F. Segovia, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, I. Alvarez, M. Lopez, and R. Chaves. A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database. Neurocomputing, 75:64–71, 2012.
- [SHM+10] J. Schmidt, A. Hapfelmeier, M. Müller, R. Perneczky, A. Drzezga, A. Kurz, and S. Kramer. Interpreting PET scans by structured patient data: A data mining case study in dementia research. *Journal* of Knowledge and Information Systems (KAIS), 24:149–170, 2010.
- [SPSH05] G. Sigletos, G. S. Paliouras, C. D. Spyropoulos, and M. Hatzopoulos. Combining information extraction systems using voting and stacked generalization. *Journal of Machine Learning Research*, 6:1751–1782, 2005.
- [Teh11] Y. W. Teh. Dirichlet process. Encyclopedia of Machine Learning, pages 280–287, 2011.
- [Tha08] N. A. Thacker. Tutorial: A critical analysis of voxel based morphometry (VBM). *University of Manchester*, 2008.
- [TMLP+02] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello,
 O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated ana-

tomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289, 2002.

- [TW97] K. M. Ting and I. H. Witten. Stacking bagged and dagged models. pages 367–375, 1997.
- [TW99] K. M. Ting and I. H. Witten. Issues in stacked generalization.

 Journal of Artificial Intelligence Research, 10:271–289, 1999.
- [VGVZ09] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the International Conference on Computer Vision*, 2009.
- [vLK11] M. V. van Leeuwen and A. Knobbe. Non-redundant subgroup discovery in large and complex data. In *Proceedings of the 21st European Conference on Machine learning and Principles and Practice of Knowledge Discovery in Databases*, pages 459–474. Springer, 2011.
- [VLS11] J. Vreeken, M. Leeuwen, and A. Siebes. Krimp: mining itemsets that compress. *Data Mining Knowledge Discovery*, 23:169–214, 2011.
- [Wan09] X. R. Wang. Structured Topic Models: Jointly Modeling Words and Their Accompanying Modalities. PhD thesis, University of Massachusetts Amherst, 2009.
- [WFH11] I. H. Witten, E. Frank, and M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques: Third Edition. The Morgan Kaufmann Series in Data Management Systems, 2011.
- [Wol92] D. H. Wolpert. Stacked generalization. Neural Networks, 5(2):241–259, 1992.
- [Wro97] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, 1997.

[WWS12] S. Weintraub, A. H. Wicklund, and D. P. Salmon. The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, 2(4):a006171, 2012.

- [WYC06] S. Q. Wang, J. Yang, and K. C. Chou. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *Journal of Theoretical Biology*, 242:941–946, 2006.
- [XCYH06] D. Xin, H. Cheng, X. F. Yan, and J. W. Han. Extracting redundancy-aware top-k patterns. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 444–453, 2006.
- [YLB⁺08] I. Yakushev, C. Landvogt, H. G. Buchholz, A. Fellgiebel, Hammers. A., A. Scheurich, I. Schmidtmann, A. Gerhard, M. Schreckenberger, and P. Bartenstein. Choice of reference area in studies of Alzheimer's disease using positron emission tomography with fluorodeoxyglucose-F18. *Psychiatry Research: Neuroimaging*, 164(2):143–153, 2008.
- [YLTS14] G. Yu, Y. Liu, K. H. Thung, and D. Shen. Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals. *PLoS ONE*, 9(5):735–745, 2014.
- [YMC+13] J. Young, M. Madat, M. J. Cardoso, A. Mendelson, D. Cash, and S. Ourselin. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical, 19(2):735–745, 2013.
- [ZLXJ11] Z. W. Zhai, B. Liu, H. Xu, and P. F. Jia. Constrained LDA for grouping product features in opinion mining. In Proceedings of the 15th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part I, pages 448–459, 2011.
- [ZR04] A. Zimmermann and L. De Raedt. CorClass: Correlated association rule mining for classification. In *Proceedings of Discovery Science*, pages 60–72. Springer, 2004.

[ZWZ+11] D. Q. Zhang, Y. P. Wang, L. P. Zhou, H. Yuan, and D. G. Shen. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage, 55(3):856–867, 2011.