Technical Note PR-TN 2008/00335

Issued: 06/2008

OPTIMAL FEATURE EXTRACTION FOR THE CLASSIFICATION OF MEDICAL IMAGES

A. Serag; F. Wenzel; F.O. Thiele; S. Young

Philips Research North-America,

Philips Research Europe

Unclassified

© Koninklijke Philips Electronics N.V. 2008

Authors' address

A. Serag ahmed.serag@philips.com
F. Wenzel fabian.wenzel@philips.com
F.O. Thiele frank.o.thiele@philips.com
S. Young stewart.young@philips.com

© KONINKLIJKE PHILIPS ELECTRONICS NV 2008
All rights reserved. Reproduction or dissemination in whole or in part is prohibited without the prior written consent of the copyright holder .

Title: OPTIMAL FEATURE EXTRACTION FOR THE CLASSIFICATION OF MEDICAL IMAGES

Author(s): A. Serag; F. Wenzel; F.O. Thiele; S. Young

Reviewer(s): Meiling Schmelzer

Technical Note: PR-TN 2008/00335

Additional Numbers:

Subcategory:

Project: Computer Aided Diagnosis for Dementia (2006-227)

Customer:

Keywords: classification accuracy, PET, brain tissue classification, automatic classification, alzheimer disease, neuroimaging, neurodegenerative diseases, Alzheimer's disease

Abstract: Dementia is significant loss of intellectual abilities such as memory capacity, severe enough to interfere with social or occupational functioning. The most common types of dementia are: Alzheimer's disease (AD), Lewy body dementia (LBD), and frontotemporal dementia (FTD). In 2008, there are currently 29.8 million people with dementia, with the number expected to be 81.1 million by 2050.

The classification of FDG-PET in patients with dementia might not be an easy task. Some dementia diseases have similar disease patterns which lead to the misdiagnosis of these diseases. Nowadays, images are visually evaluated by an expert reader and this process is not entirely quantitative or reproducible. Even the experts can have up to 20% misclassification. Automated diagnosis by pattern recognition can produce quantitative and reproducible results, but if training data comes from clinical routine, it may produce less accurate results to discriminate similar disease patterns. Hence, the work of this thesis aims to find an optimal subset of features, for a given training data set, which improves the classification of FDG-PET in patients with suspected dementia.

© Koninklijke Philips Electronics N.V. 2008

The presented work describes some methods that found effective in improving the classification problem at hand. These methods consist of: (i) feature extraction, (ii) feature ranking, (iii) classifier learning with balanced data, (iv) feature selection. The results demonstrate the possibility to improve the accuracy of pair-wise classification for the most common dementia diseases with an excellent accuracy (in less time) by these methods.

 © Koninklijke Philips Electronics N.V. 2008

OPTIMAL FEATURE EXTRACTION FOR THE CLASSIFICATION OF MEDICAL IMAGES

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science in Biomedical Engineering

Presented to

Lübeck University of Applied Sciences

&

University of Lübeck

by

Ahmed SERAG

Carried out at

Philips Research Hamburg

Under the supervision of

Examiner: Prof. Dr. rer. nat. Thorsten M. BUZUG (University of Lübeck)

Co-examiner: Prof. Dr.-Ing. Alfred MERTINS (University of Lübeck)

External supervisor: Dr. Stewart YOUNG (Philips Research Hamburg)

External supervisor: Dr.-Ing. Fabian WENZEL (Philips Research Hamburg)

2008

Universität zu Lübeck

PHILIPS
sense and simplicity

FACH HOCHSCHULE LÜBECK
University of Applied Sciences

ABSTRACT

Dementia is significant loss of intellectual abilities such as memory capacity, severe enough to interfere with social or occupational functioning. The most common types of dementia are: Alzheimer's disease (AD), Lewy body dementia (LBD), and frontotemporal dementia (FTD). In 2008, there are currently 29.8 million people with dementia, with the number expected to be 81.1 million by 2050 [1].

The classification of FDG-PET in patients with dementia might not be an easy task. Some dementia diseases have similar disease patterns which lead to the misdiagnosis of these diseases. Nowadays, images are visually evaluated by an expert reader and this process is not entirely quantitative or reproducible. Even the experts can have up to 20% misclassification. Automated diagnosis by pattern recognition can produce quantitative and reproducible results, but if training data comes from clinical routine, it may produce less accurate results to discriminate similar disease patterns. Hence, the work of this thesis aims to find an optimal subset of features, for a given training data set, which improves the classification of FDG-PET in patients with suspected dementia.

The presented work describes some methods that found effective in improving the classification problem at hand. These methods consist of: (i) feature extraction, (ii) feature ranking, (iii) classifier learning with balanced data, (iv) feature selection. The results demonstrate the possibility to improve the accuracy of pair-wise classification for the most common dementia diseases with an excellent accuracy (in less time) by these methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor, *Dr. Stewart Young*, for his support and encouragement for me. I am greatly indebted to him for the many occasions in which he has gone out of his way to assist in the timely completion of this work.

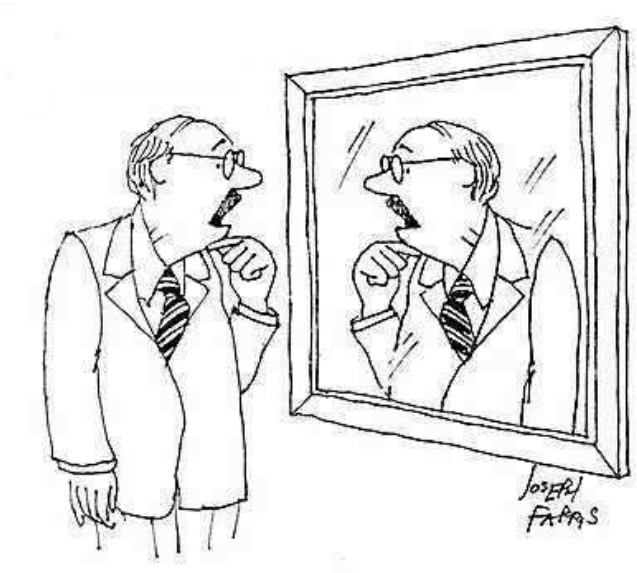
This research would not have been possible without the support of *Dr.-Ing. Fabian Wenzel* who was abundantly helpful and offered invaluable assistance, support and guidance.

Special thanks go to *Dr. rer. nat. Lothar Spies* for motivating my work. I am extremely grateful to him for his support as *Director of the Digital Imaging Department* at *Philips Research Hamburg*.

I feel that the environment in the *Digital Imaging Department* here at *Philips Research Hamburg*, has contributed greatly to this work. Whenever a problem arose, I always felt that I could go and ask practically anyone about it, and if they didn't know the answer, they could point me towards someone who did. I have a feeling that it is this sort of environment that I will miss most when I will leave for another challenge.

Special thanks also to *Dr. rer. nat. Ralph Buchert* without whose knowledge and assistance this study would not have been successful.

Deepest gratitude is due to the members of the examination committee at *Lübeck University*, *Prof. Dr. rer. nat. Thorsten M. Buzug* and *Prof. Dr.-Ing. Alfred Mertins*.



"I remember the face but I've forgotten your name. "

C h a p t e r 1

INTRODUCTION

Summary

C h a p t e r 1

INTRODUCTION

1.1 Overview

The classification of FDG-PET in patients with dementia might not be an easy task. Some dementia diseases have similar disease patterns which lead to the misdiagnosis of these diseases. Nowadays, images are visually evaluated by an expert reader and this process is not entirely quantitative or reproducible. Even the experts can have up to 20% misclassification. Automated diagnosis by pattern recognition can produce quantitative and reproducible results, but if training data comes from clinical routine, it may produce less accurate results to discriminate similar disease patterns. Hence, the work of this thesis aims to find an optimal subset of features, for a given training data set, which improves the classification of FDG-PET in patients with suspected dementia.

1.2 About Dementia

Dementia is significant loss of intellectual abilities such as memory capacity, severe enough to interfere with social or occupational functioning[1]. It is a broad term used to describe a loss of memory, intellect, rationality, social skills and what would be considered normal emotional reactions.

Although most of the dementia patients are old, not all old people get dementia. This means that dementia is not a normal part of ageing. Dementia can happen to any person, but it is more common after reaching the age of 65 years. Persons in 40s and 50s can also have dementia. Both men and women can develop dementia, although it is more common in men [1].

In 2008, there are currently 29.8 million people with dementia, with the number expected to be 81.1 million by 2050. It is estimated there will be 4.6 million new cases of dementia every year (one new case every 7 seconds). The number of people affected will double every 20 years to 81.1 million by 2040 [2, 3].

[1] Definition of dementia: http://webmed.com/mental-health/dementia

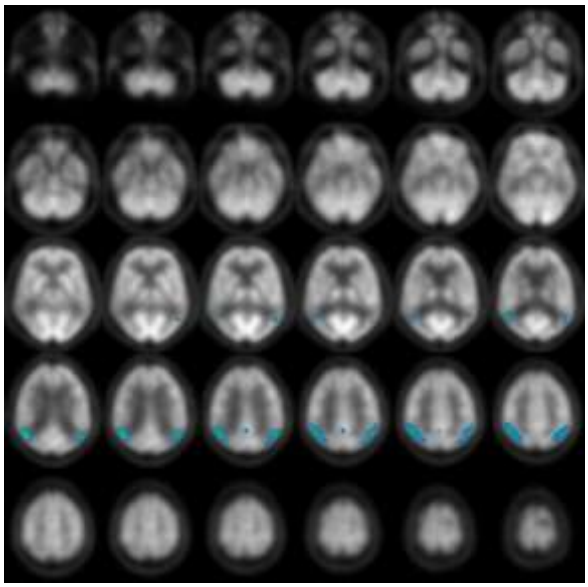


Fig. 1. 1 - FDG-PET scan of a patient with Alzheimer's disease

1.3 Types of Dementia

There are many different types of dementia and each has its own causes. Some of the most common types of dementia are:

Alzheimer's disease (AD)

Alzheimer's disease is the most common cause of dementia. It accounts for between 50% and 70% of all dementia cases. It is a progressive, degenerative illness that attacks the brain and causes shrinkage or disappearance of brain cells. Then, abnormal material builds up as "tangles" in the centre of the brain cells, and "plaques" outside the brain cells. These interrupt messages within the brain and destroy connections between brain cells. Sooner or later, the brain cells die and information cannot be recalled or taken in anymore. As Alzheimer's disease affects neurons in the brain, certain functions or abilities start to be affected. At the beginning, memory of recent events is affected, but as the disease progresses, long-term memory is also affected. Alzheimer's also affects many of the brain's other functions and consequently many other aspects of behaviour are affected [4].

In the early stages, the symptoms of Alzheimer's disease can be very slight; however, it often begins with lapses of memory and difficulty in finding the right words for everyday objects.

Nowadays, researchers are learning more about the physiological changes that destroy brain cells in Alzheimer's disease, but apart from the few individuals with familial Alzheimer's disease, it is not known why one individual gets Alzheimer's disease and another does not. A variety of suspected causes are being investigated, including factors in the environment, biochemical disturbances and immune processes. The cause may vary from person to person and may be due to one factor or a number of factors.

Lewy body dementia (LBD)

Lewy body dementia is caused by the degeneration and death of nerve cells in the brain and it accounts for between 10% and 15% of all dementia cases. The name comes from the presence of abnormal spherical structures, called Lewy bodies, which develop inside nerve cells. It is thought that these may lead to the death of the brain cells. People who have Lewy body dementia tend to see things (visual hallucinations), experience stiffness or shakiness, and their condition tends to fluctuate quite rapidly, often from hour to hour or day to day. Although these symptoms allow it to be differentiated from Alzheimer's disease, Lewy body dementia is similar to Alzheimer's disease in many ways and sometimes Lewy body dementia co-occurs with Alzheimer's disease [5].

The symptoms of Lewy body dementia include difficulty with concentration and attention, extreme confusion and difficulty judging distances (often results in falls). Some people who have also Lewy body dementia may also experience delusions and/or depression.

At present there is no known cause of Lewy body dementia and no known risk factors have been identified. There is no evidence that Lewy body dementia is an inherited disease.

Frontotemporal dementia (FTD)

Frontotemporal Lobar Degeneration (FTLD) is the name given to dementia when there is degeneration in one or both of the frontal or temporal lobes of the brain. Frontotemporal dementia (FTD) is the most common subtype. It accounts for about 10% of all dementia cases. It is mainly a disorder of behaviour. People with Frontotemporal dementia may be disinhibited or apathetic [6].

Early symptoms of Frontotemporal dementia can affect behaviour, and sometimes language. People may show change in their character and in their social behaviour. A person with Frontotemporal dementia may become obsessive

and repeat the same action over and over again. Language problems often occur early in the disease and may range from limited speech to total loss of speech.

About 50% of people with Frontotemporal dementia have a family history of the disease. Those who inherit it often have a mutation in the tau protein gene on chromosome 17 leading to abnormal tau protein being produced. No other risk factors are known.

1.4 The Need for Early Diagnosis

Alzheimer's and other forms of dementia will have a major impact on future medical care, caused by mainly three reasons:

(i) The increasing life expectancy
(ii) Lack of cures
(iii) High cost of available cures

Ongoing research in many sites demonstrates the urgent need for reliable early and differential diagnosis, which can ensure that a patient gets the right treatment at the right time – currently available medications aim towards alleviating the symptoms, but in the future medicines may become available which treat the underlying cause.

In contrast, early diagnosis enables persons with dementia and their family to receive help in understanding and adjusting to the diagnosis and to prepare for the future in an effective way. This might include making legal and financial actions, changes to living activities, and learn about services that can improve quality of life for people with dementia and their friends and family. It also can help families to educate themselves about the disease and learn effective ways of interacting with the person who has dementia [7]. There is also evidence that AD treatments are more effective when given earlier in the disease process.

1.5 Dementia Diagnosis by Imaging Techniques

Diagnosing dementia is done in many ways today. Beside psychiatric tests, like the minimal state examination (MMSE), blood and spinal fluid tests are able to grade and identify different neurological diseases. However, as each type of dementia can be characterized by reduced activity in the brain, imaging plays an important role.

Conventional CT and MR Imaging

Conventional structural neuroimaging, such as computed tomography (CT) or magnetic resonance (MR), are used for the routine diagnosis of dementia. These are recommended in diagnostic guidelines for AD, principally in order to exclude other potential causes of the symptoms. Nevertheless, the low sensitivity and specificity for the diagnosis makes it only to be used as an add-on to help weigh up the degree of atrophy and rule out other causes of dementia (such as normal-pressure hydrocephalus, vascular dementia, or intracranial mass) [8].

With conventional neuroimaging modalities, sensitivity about 85% may be obtained for patients with dementia [8]. Nevertheless, as structural changes occur late in the course of the disease, conventional structural neuroimaging (CT and MR) may not be the best choice in identifying slight pathologic changes early enough throughout the disease course.

Functional Neuroimaging

Functional imaging techniques, such as single photon emission computed tomography (SPECT), positron emission tomography (PET), or functional magnetic resonance imaging (fMRI), offer a potentially valuable alternative in the differential diagnosis of dementia, mainly in distinguishing Alzheimer's disease from Frontotemporal dementia and Lewy body dementia. Furthermore, functional imaging modalities, including SPECT, PET, and fMRI, could have greater chance in identifying more slight pathologic changes earlier in the disease course.

SPECT was reported to be able to distinguish AD patients from healthy control subjects with a high degree of sensitivity around 89% [8]. Furthermore, the accuracy of a clinical diagnosis of AD is improved with the assist of this technique. Nevertheless, adding SPECT imaging to the diagnostic course is not cost-effective given currently existing therapies.

PET studies have been successful in discriminating AD from other forms of dementia on the basis of the pattern of FDG uptake. The majority of PET studies in the memory impairment population have used FDG as a radiolabeled tracer which was able to provide overall sensitivity for detecting AD with sensitivity of 94% (in patients with probable AD) [8]. Hence, FDG-PET is a good choice in the discrimination between different forms of dementia.

As there are always two sides to every story, Tab. 1.1 summarizes the relative merits for CT, MR, SPECT and PET. We can conclude through the strengths and weaknesses of these imaging techniques that functional

neuroimaging are doing a better job in dementia diagnosing. In the remainder of this report, only FDG-PET images are discussed.

Modality	Strengths	Weaknesses
CT	- excellent spatial resolution - relatively cheap - widely available - rules out major pathologies	- less contrast - problem of bone artifact - less sensitive to pathology - patient is exposed to radiation
MR	- excellent spatial resolution - no bone artifact - better contrast with high sensitivity - low risk	- relatively higher cost - less readily available - longer imaging time (this is changing) - associated considerations (i.e. pacemaker)
SPECT	- relatively cheap and widely available - injected radiolabel measures regional brain perfusion	- only offers relative quantification - poor spatial resolution - has a risk of radiation exposure
PET	- fair resolution - direct quantification is possible - brain activation can be measured using subtraction	- high costs associated (cyclotron and special tem) - scarce resource and not widely available - a risk associated with exposure to radiation

Tab. 1. 1 - Strengths and weaknesses of imaging techniques for diagnosis of dementia

1.6 Computer-Aided Evaluation of Dementia

First of all, it's important to mention that there's no definitive test for diagnosing dementia. Findings from tests and a diversity of sources have to be collected before diagnosis can be done and this process can be complex and time-intensive. Furthermore, a degree of uncertainty remains, as for example indicated by the diagnostic classes typically used ("possible" or "probable"). The only 100% reliable approach remains pathological analysis after death.

Without computer-aided diagnosis, detection of hypo-metabolic areas in FDG-PET has to be done by an expert in nuclear medicine and requires knowledge of typical variations in PET brain images as well as distribution of areas in the brain, i.e. patterns, that indicate specific neurological diseases. Manual interpretation remains subjective, not reproducible, and cannot be done fully quantitatively.

The Philips Research Europe – Hamburg project "Computer-aided Evaluation of Dementia" started in 2005. Features of PET-based research include:

(i) Highlighting significantly hypo-metabolic regions in the scan
(ii) The automatic interpretation of PET scans in the sense that likelihoods for specific types of dementia are estimated

Ultimately, the project aims to help an expert as an objective, quantitative, second reader without any intention to push away physicians [9].

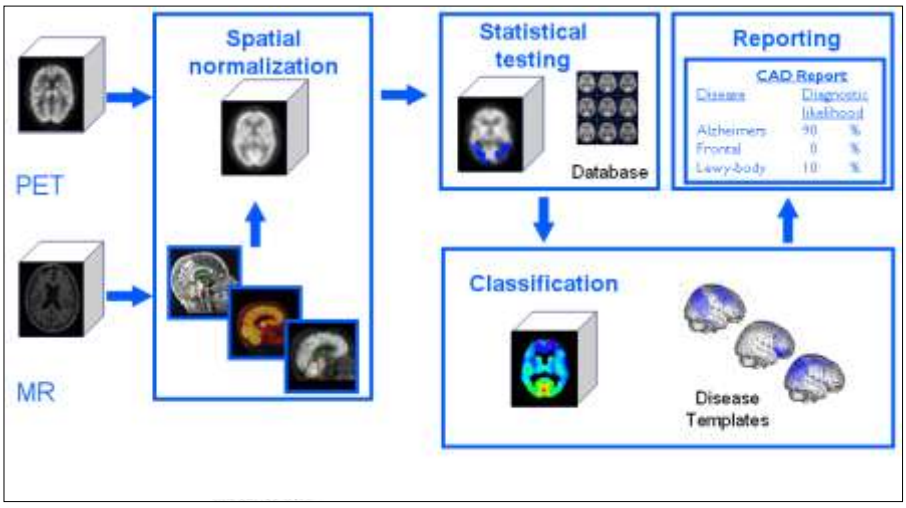


Fig. 1. 2 - An overview of an approach to differential diagnosis with FDG-PET as explored at Philips Research

Fig. 1.2 shows an overview of the project Computer-Aided Evaluation of Dementia at Philips Research. Two key methods are used, firstly elastic registration whereby images are aligned automatically to a reference brain. Secondly, a comparison of the aligned images to typical disease patterns, stored in a database, is made.

1.7 Automated Classification of FDG-PET Images for Dementia Patients

FDG-PET brain scans were acquired on 92 subjects as in Tab. 1.2.

Group	N	Age (y)	% Female
AD	44	64.3 ± 7.0	40 %
LBD	9	69.6 ± 8.0	22 %
FTD	13	59.5 ± 9.7	77 %
NC	26	49.6 ± 13.2	53%

Tab. 1. 2 - Demographics of subjects used in the differential diagnosis.
Age is given as mean ± standard deviation of the sample

All FDG-PET images were aligned to a customized FDG-PET template with a voxel size of 2 x 2 x 2 mm^3 [10]. Images were subsequently smoothed using an isotropic Gaussian filter with 10 mm FWHM, and finally scaled to a common median intensity value inside a pre-defined gray-and white-matter mask.

The resulting images were classified using a linear regression-based approach. System performance was tested in a leave-one-out cross-validation (LOOCV) for pair-wise classification problem (all possible 2-class combinations), and for the multi-class (4-class) classification problem, whereby each class was trained using a one-versus-all approach. Isotonic regression was applied to convert classifier scores to likelihoods, and a maximum-likelihood decision rule applied to assign class labels. A receiver operating characteristic (ROC)[2] analysis was performed to compare results and area under ROC curve (AUC) was used to measure the classification accuracy.

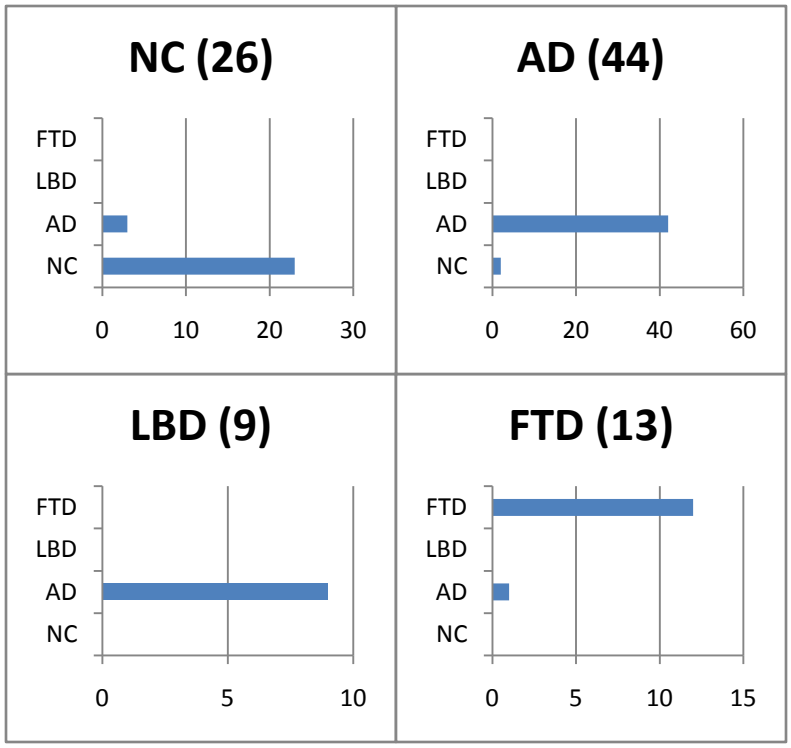


Fig. 1. 3 - Result of voxel-based automated classification of FDG-PET images of dementia patients

[2] ROC curve is a graphical plot of the sensitivity vs. (1-specificity) for a binary classifier system as its discrimination threshold is varied. Area under ROC curve (AUC) is a very useful measure of similarity between two classes [27].

Fig 1.3 shows that voxel-based automated classification has some misclassification errors. The most notable misclassification can be seen clearly in discriminating LBD from other forms of dementia. Even when trying to simplify the discrimination problem to 2-class problems, still the classifier cannot discriminate LBD from AD which is the most critical case (See Fig. 1.4).

Discriminating LBD from AD is not an easy task due to several reasons. Regarding the clinical symptoms, both diseases share similar symptoms and they can even potentially occur together at the same time (See section 1.3). What makes it worse that the LBD database contains, only, 9 patients while the AD database contains 44 patients. This large difference between both databases makes it difficult to train well our classifier in order to be able to differentiate automatically between both diseases.

Working on this special case will help not only to solve the AD-LBD discrimination issue, but also to have insight into the problem at hand which could help to deal with future problems with similar issues, i.e. similar disease patterns and/or notable difference between the sizes of diseases' databases.

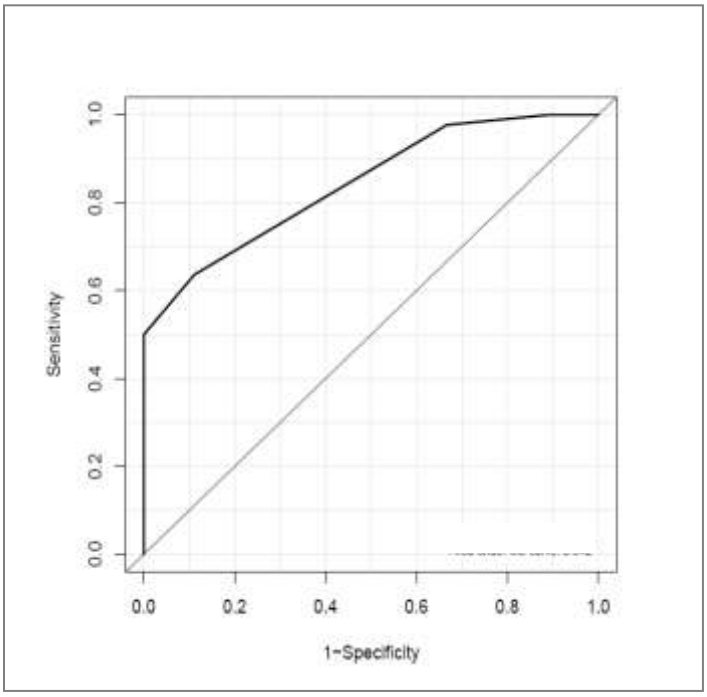


Fig. 1. 4 - ROC Curve for the result of AD-LBD discrimination (AUC = 0.842)

In *chapter 2 "Feature Extraction",* the extraction of the relevant features using a brain atlas is described. The purpose of feature extraction is to reduce the amount of data which is to be analyzed in subsequent steps. It also increases robustness of a classifier as the influence of random variations in single voxel intensities is reduced.

In *chapter 3 "Feature Ranking",* some methods to rank the relevant features are described. Although feature ranking is not required to build classifiers, several feature selection algorithms take account of feature ranking as a chief or supplementary selection mechanism due to its simplicity, good empirical success and scalability.

In *chapter 4 "The Influence of Class Imbalance on Training"*, the class imbalance and its influence on the classifier training is studied. Although most of today's research on classification training has focused on the training algorithm itself, some studies have demonstrated that the class distribution has a significant effect on the training process [11, 12].

In *chapter 5 "Feature Selection",* several methods to select the optimal features are described. The most common feature selections algorithms, as well as a proposed one, are evaluated. Here, there is no attempt to generate new features, only selection of $d < D$ features is done by getting rid of irrelevant features and maintaining the relevant ones. That will result in having insight into the nature of the classification problem, improving of the classification accuracy, and simplifying the construction of the classifier.

In *chapter 6 "Conclusion & Future Work",* the work done is summarized, and future work is mentioned.

C h a p t e r 2

FEATURE EXTRACTION

Summary

C h a p t e r 2

FEATURE EXTRACTION

2.1 Introduction

A potential obstacle in the application of pattern recognition arises from the high dimensionality of the input data. Dimension reduction is one approach for removing this obstacle. Simply, dimension reduction is the process of reducing the number of input data without a significant loss in information.

Dimension reduction plays an important role in classification processes by transforming the data into another representation (called feature vectors). Sometimes, the classification can be done more accurately in the reduced space than in the original one.

Dimension reduction can be divided into: *feature extraction* and *feature selection*. F*eature extraction* is the process of generating the features from the input data and *feature selection* is the process of selecting the best features form the current set of features. In this chapter *feature extraction* is discussed and in *chapter 5* the other half of the dimension reduction process, *feature selection*, is explained.

In the described setup of the studied system, the size of a 3D scan is 91 x 109 x 91 voxels (1.7 x 3.4 x 1.7 mm^3). Clinically relevant parts of the image cover gray and white matter of about 500000 voxels. For training a classifier with many scans, this amount of data might already be a computational problem (i.e. ~100 images consisting of 500000 voxels for each training). However, due to the low resolution of PET images and smoothing during pre-processing steps, adjacent voxels do not contain independent information. There is a spatial coherence of information between a voxel and its neighbors. So, it is not necessary to consider each voxel for automatic classification, rather it is possible to focus on lower-resolution information [9].

One possibility is to combine voxel intensities in regions of the brain with high functional correlation. Not only does such a feature extraction reduce the amount of data which is to be analyzed in subsequent steps, it also increases robustness of a classifier as the influence of random variations in single voxel intensities is reduced.

As feature extraction is part of the dimension reduction, in a typical classification task, if the number of relevant features (voxels) is N, the feature extraction problem is defined as obtaining the $n < N$ features that enable the construction of the best classifier.

Now, the question is how to extract these features? There are many ways to extract features. Here, features are extracted using "brain atlas". The idea behind the brain atlas is not new. The main concept is to identify and target specific regions in the brain.

A wide-range of maps of the brain structure already exist. Recently atlases and nearly all early ones were derived from one or at best a few individual post-mortems specimens. Nowadays, brain atlases are being developed in such a way as to include flexible, computable systems, which describe the most significant variation in a population [13].

The features are extracted by masking the pre-processed PET images with the brain mask. This leads to the extraction of the anatomical volumes of interest (AVOI). Then, each AVOI is represented by the mean value of the intensities inside this AVOI. At the end, each image will be represented by a feature vector $F = (f_1, f_2, \dots, f_n)$ where n is the number of AVOIs.

Fig. 2. 1 - Feature extraction using brain atlas

In the following sections of this chapter, we discuss the results of using two different brain atlases, and which is better with respect to the classification task at hand:

(i) Anatomical Automatic Labeling (AAL) Atlas
(ii) Universitätsklinikum Hamburg-Eppendorf (UKE) Atlas

2.2 Anatomical Automatic Labeling (AAL) Atlas

The anatomical automatic labeling (AAL) atlas is a brain atlas volume which is included in the AAL software package which designed to be used with the SPM analysis package[3]. The AAL atlas identifies anatomical regions on a brain that is in the MNI[4] space [14] The MNI wanted to describe a brain that is more representative of the population. First, they used 250 normal MRI scans to build the 250 atlas brain which is seldom used. Second, an extra 55 images were made and registered to the 250 atlas using an automatic linear registration method. Then, the registered 55 brains were averaged with the 250 manually registered brains to create the MNI 305 atlas [15].

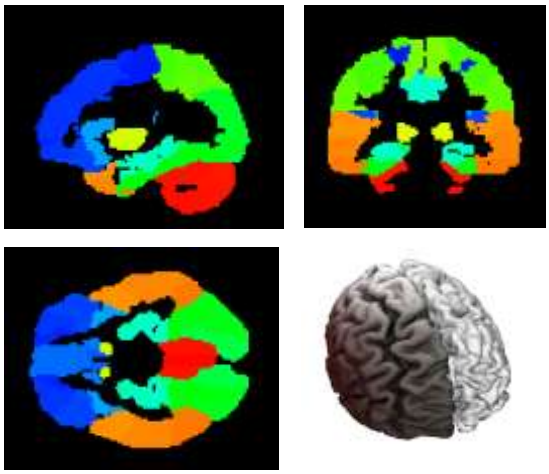


Fig. 2. 2 - The AAL brain atlas

The AAL alas consists of 116 regions of interest (ROIs). This means a significant dimension reduction of the input data (comparing to ~500000 voxels). Using the AAL atlas image volume as a mask, one could then restrict the processing to particular brain regions, or, identify ROIs that one computes with the whole brain analysis.

[3] The SPM (Statistical Parametric Mapping) is a software package has designed for the analysis of brain imaging data sequences. The sequences can be a series of images from different cohorts, or time-series from the same subject. The current release (SPM5) is designed for the analysis of fMRI, PET, SPECT, EEG and MEG. Refer to: http://www.fil.ion.ucl.ac.uk/spm

[4] SPM 96 and later uses standard brains from the Montreal Neurological Institute. The MNI defined a new standard brain by using a large series of MRI scans on normal controls.

Region Anatomical Description
Precentral
Frontal_Sup
Frontal_Sup_Orb
Frontal_Mid
Frontal_Mid_Orb
Frontal_Inf_Oper
Frontal_Inf_Tri
Frontal_Inf_Orb
Rolandic_Oper
Supp_Motor_Area
Olfactory
Frontal_Sup_Medial
Frontal_Med_Orb
Rectus
Insula
Cingulum_Ant
Cingulum_Mid
Cingulum_Post
Hippocampus
ParaHippocampal
Amygdala
Calcarine
Cuneus
Lingual
Occipital_Sup
Occipital_Mid
Occipital_Inf
Fusiform
Postcentral
Parietal_Sup
Parietal_Inf
SupraMarginal
Angular
Precuneus
Paracentral_Lobule
Caudate
Putamen
Pallidum
Thalamus
Heschl
Temporal_Sup
Temporal_Pole_Sup
Temporal_Mid
Temporal_Pole_Mid
Temporal_Inf
Cerebelum_Crus1
Cerebelum_Crus2
Cerebelum_3
Cerebelum_4_5
Cerebelum_6
Cerebelum_7b
Cerebelum_8
Cerebelum_9
Cerebelum_10
Vermis_1_2
Vermis_3
Vermis_4_5
Vermis_6_7_8_9_10

Tab. 2. 1 - List of the AAL atlas ROIs for each hemisphere

2.3 Universitätsklinikum Hamburg-Eppendorf (UKE) Atlas

The AAL-Atlas works but not optimized for PET. The UKE atlas is an optimized version of the AAL, which is made by University Medical Center Hamburg-Eppendorf. The aim was to create a digital brain atlas for SPECT and PET, in particular, for the typical finding patterns in neurodegenerative diseases [16]. That UKE atlas consists of 29 AVOIs, instead of 116 AVOIs in the AAL atlas. This means a dimension reduction of 75% that can give more insight into the nature of the problem at hand. Also, this reduction leads to efficient computation time.

Region Anatomical Description
Fusiform
Prefrontal_mesial
Rectus
Temporal_ pole
Pons
Temporal_ mesial
Temporal_lat
Precuneus
Parietal_sup
Parietal_inf
Insula
Heschle
Orbito_frontal
Occipital
Prefrontal_lat
Cerebellum
Supp_motor_cortex
Motor_cortex
Cingulum_post
Cingulum_mid
Cungulum_ant
Sensory_cortex
Lingual_wo_brod_71
Thalamus
Caudatus
Medulla_oblongata
Lentiforme
Midbrain
Brodmann17

Tab. 2. 2 - The UKE atlas ROIs

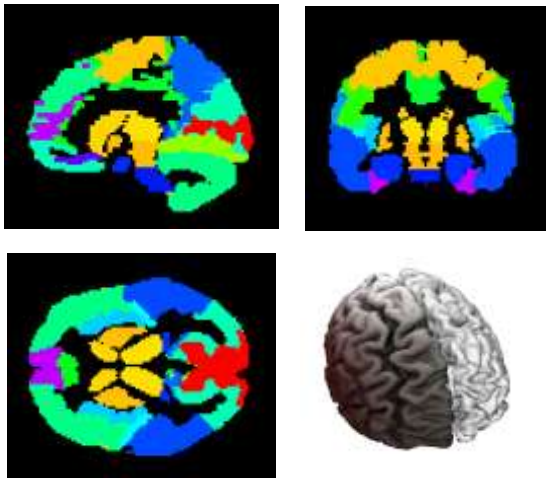


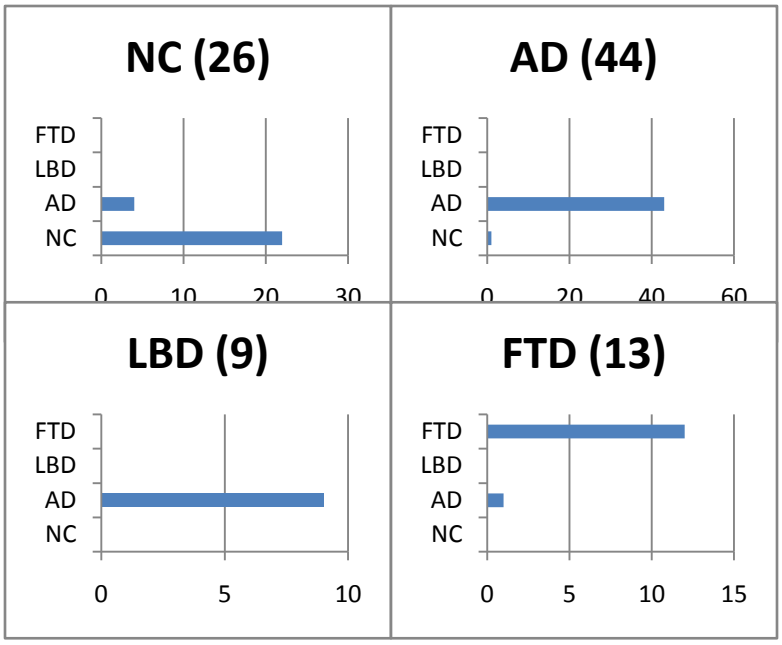
Fig. 2. 3 - The UKE brain atlas

2.4 The AAL Atlas vs. The UKE Atlas

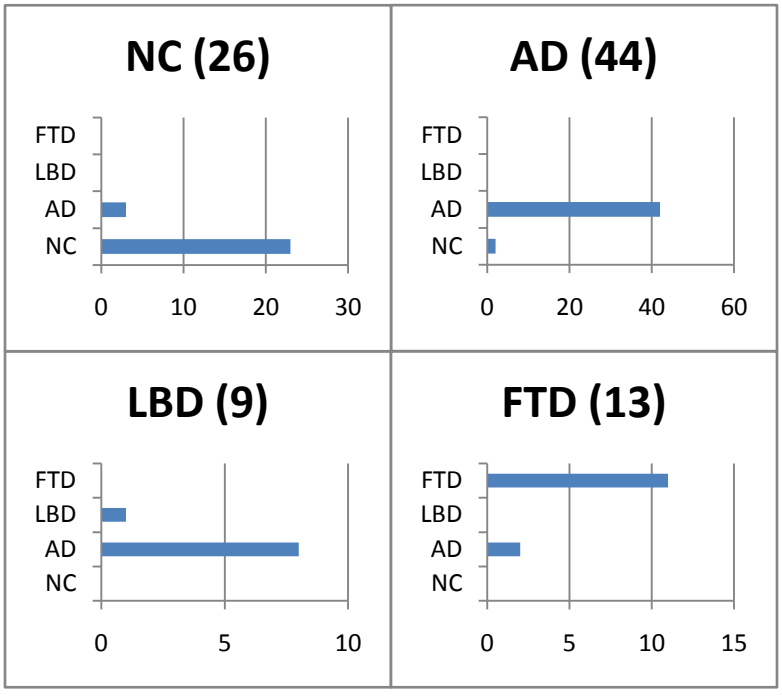
Although the UKE atlas consists of only 29 ROIs while the AAL consists of 116 ROIs, the UKE atlas gives comparable results to the AAL one (See Fig. 2.4, 2.5). The classification results are almost the same for 2-class problems and 4-class problem, and hence arises a question: as long as both atlases give similar results, which one should be used?

Of course, it's cheaper to process only 29 features (UKE feature-based) than 116 features (AAL feature-based) or ~500000 features (voxel-based). This means an efficient computation time can be obtained. On the other hand, studying less number of features allows having insight into the classification problem. In other words, the influence of individual features on classifier learning can be studied (will be seen in the next chapter). Hence, the features which could be omitted in order to potentially improve classification accuracy can be found.

In addition, reducing the number of features affect later processes, like feature selection process. For instance, if the classifier itself is used to select the optimal subset of features (as in the wrapper approach), handling 116 features is not an easy job and can take many hours to get the optimal selected subset of features. It is better not to continue talking about the influence of having fewer features now as this can be felt very well in further steps.

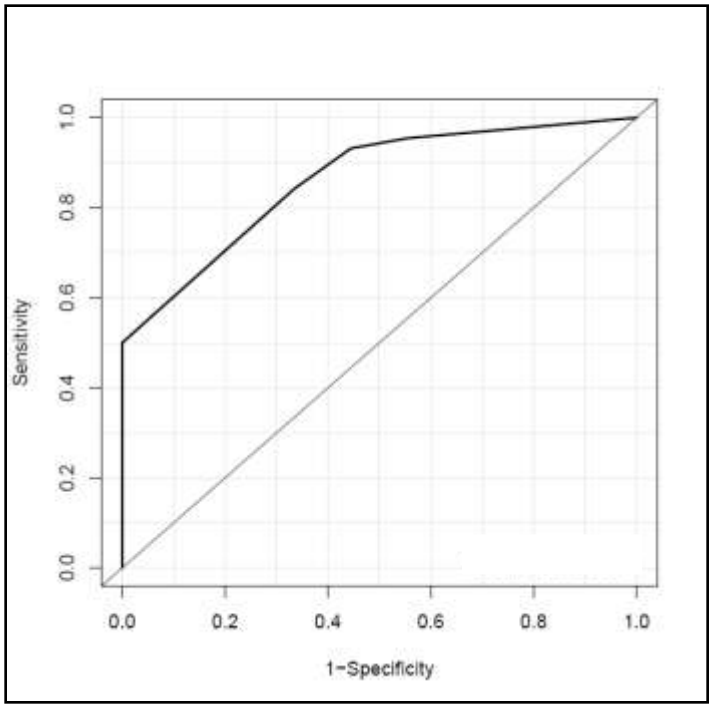


(a)

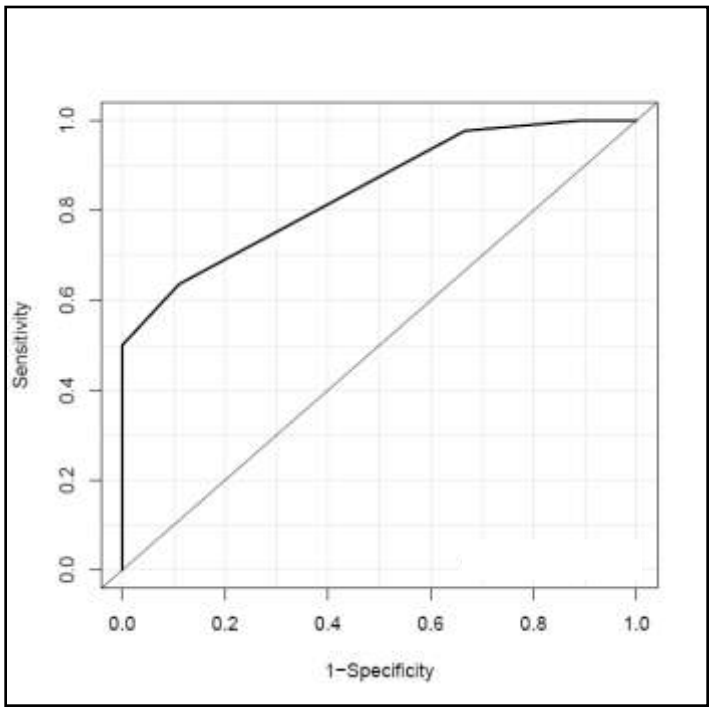


(b)

Fig. 2. 4 - Feature-based classification results using features extracted from (a) AAL atlas (b) UKE atlas



(a)



(b)

Fig. 2. 5 - ROC curves for AD-LBD feature-based discrimination using extracted features by (a) AAL atlas (AUC=0.861) (b) UKE atlas (AUC=0.842)

C h a p t e r 3

FEATURE RANKING

Summary

C h a p t e r 3

FEATURE RANKING

3.1 Introduction

Although feature ranking is not required to build classifiers, several feature selection algorithms take account of feature ranking as a chief or supplementary selection mechanism due to its simplicity, good empirical success and scalability [17]. A ranking criterion is used to find best features that discriminate between different forms of dementia.

By definition, a high score indicates a valuable feature so that features can be sorted in decreasing order of scores. Feature ranking is independent of the choice of the classifier. Additionally, even when feature ranking is not optimal, it may be preferable to other feature subset selection methods because:

(i) It is computationally efficient as it requires only computation of n scores, for $n < N$ features, and sorting these scores.
(ii) It is robust against over-fitting because it introduces bias but it may have considerable less variance [17].

In order to be able to follow the rest of this chapter, a few notations are first introduced. Consider each data to be classified is represented as a row vector x. When using a complete 3D scan with M voxels, this may be achieved by concatenating all rows and slices. All available N data sets used for training a classifier may be represented as a matrix. It is usually denoted $X = (x_1, x_2, \dots, x_n)^T$ and called data matrix. In the present context, each row x_i of X represents the features of a single scan.

As supervised learning is used, each scan belongs to a particular set of K diseases, e.g. denoted by +1 for class membership and -1 otherwise. A row vector $Y = (y_1, y_2, \dots, y_k)^T$ may be given for each x containing ± 1 in each row. As long as the correlation coefficient can only be calculated for two classes, the classification problem will be limited to pair-wise coupling and/or one-versus-all problem.

In the following sections of this chapter, correlation-based ranking and single feature classifiers are discussed.

3.2 Correlation-based Ranking (CBR)

First, let's consider the prediction of a disease (dementia type) y. The Pearson's correlation coefficient is defined as:

$$\rho_i = \frac{\sum_j \left(x_i^j - \bar{x}_i\right)\left(y^j - \bar{y}\right)}{\sqrt{\sum_j \left(x_i^j - \bar{x}_i\right)^2 \sum_j \left(y^j - \bar{y}\ \right)^2}} \qquad \text{(Eq. 3.1)}$$

$$P_i = \arg \min_{\left\{\rho_i^j \,\middle|\, 1 \le j \le n\right\}} \left|\rho_i^j\right| \qquad \text{(Eq. 3.2)}$$

$where\ i\colon feature\ number\ ;\ j\colon sample\ number$

In linear regression, the coefficient of determination, which is ${\rho_i}^2$, represents the fraction of the total variance around the mean value y^T that is explained by the linear relation between feature x_i and class y. Hence, using ${\rho_i}^2$ as a feature ranking criterion enforces a ranking according goodness of linear fit of individual features.

One question comes up a lot, why not using ρ_i directly to rank features instead or ${\rho_i}^2$. The answer is very simple as when using ρ_i , positively correlated features are ranked on top and then negatively correlated features are correlated at bottom. On the other hand, by using ${\rho_i}^2$, one can choose a subset of features with a given share of positively and negatively correlated features.

As correlation criteria like ρ_i can only discover linear dependencies between feature and class, a cheap way of elevating this constraint is to make a non-linear fit of the class with single features and rank according to the goodness of fit [17]. Still, there is a risk of over-fitting, so that one can alternatively think about using non-linear preprocessing, i.e. squaring, taking the square root, etc., and then using a simple correlation coefficient.

3.3 Single Feature Classifiers (SFC)

When dealing with a classification task one can think about using the classifier itself for ranking the features. In other words, one can rank the features according to their individual predictive power. Thus, the performance of a classifier built with a single feature to be used as a ranking criterion.

The feature discriminative power can be measured in terms of classifier performance. Although there are many ways to measure the performance of a classifier, in this task classifier performance was tested in a leave-one-out cross-validation (LOOCV) and the performance of the classifier can be described using the *confusion matrix*.

	Actual Positive	**Actual Negative**
Predict Positive	True Positive (TP)	False Positive (FP)
Predict Negative	False Negative (FN)	True Negative (TN)

True positive classification outcome (TP) and true negative classification outcome (TN) were counted in order to be used for the estimation of classification accuracy as:

$$accuracy = \frac{\sum TP + \sum TN}{\sum P + \sum N} \qquad \text{(Eq. 3.3)}$$

Eq. 3.3 is called, the traditional accuracy. For more efficient feature ranking, one can modify the previous accuracy definition to a conditional accuracy as:

$$conditional\ accuracy = \frac{\sum \alpha TP + \sum \alpha TN}{\sum P + \sum N} \qquad \text{(Eq. 3.4)}$$

$$\alpha = \begin{cases} 1 & \text{Likelihood} \geq \delta \\ 0 & otherwise \end{cases}$$

where TPs (and TNs) are only considered if, and only if, the likelihood is above certain threshold δ.

For example, in 2-class classification task, the likelihood could be 0.51 (disease) and 0.49 (not disease) and this will be considered true positive using the traditional accuracy definition (Eq. 3.3). This may lead to address the used feature as a good one, although the difference between both likelihoods is 0.02 which means that the discriminative power of the used feature is not so well. On the other hand, using the conditional accuracy definition (Eq. 3.4), the outcome result will be considered true positive (or true negative) if, and only if, the difference between both likelihoods is greater than defined threshold δ.

In this case if there is a large number of features which differentiate the disease classes perfectly, ranking criteria based on conditional accuracy of classification can make a distinction between the top ranking features vey well. But, in case of using the traditional accuracy definition, one would prefer to use a correlation based feature ranking.

3.4 CFR vs. SFC

Fig. 3.1 shows the average correlation coefficient for each feature according to CBR ranking.

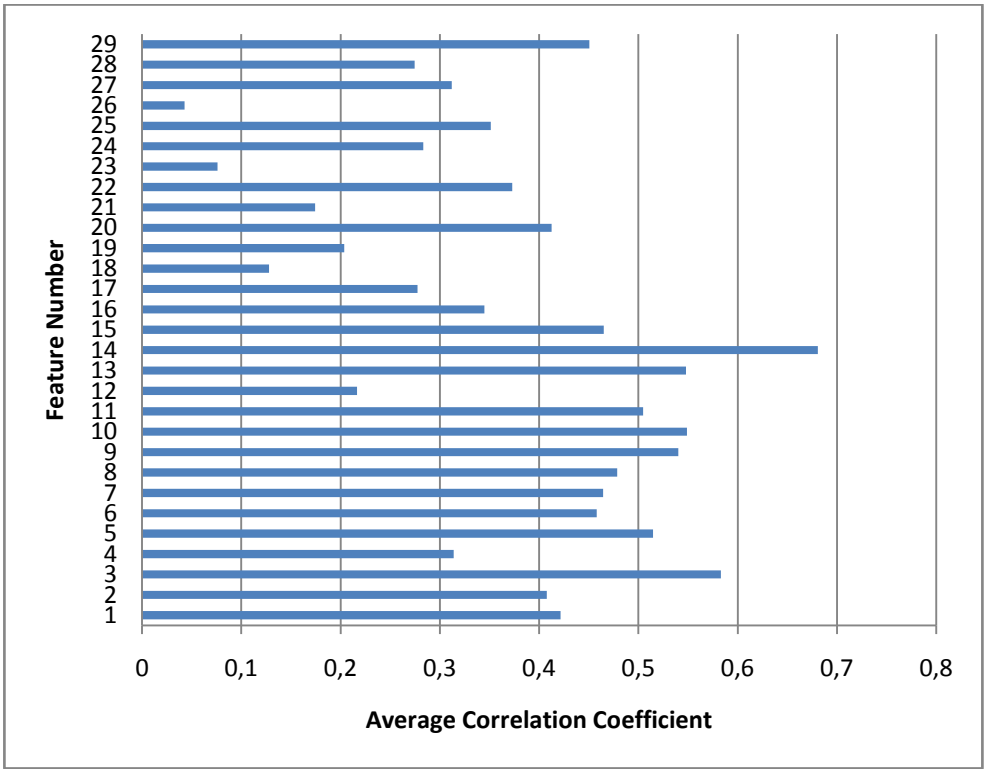


Fig. 3. 1 - Average correlation coefficient for each feature according to CBR

Fig. 3.2 shows the discrimination power for each feature according to SFC ranking.

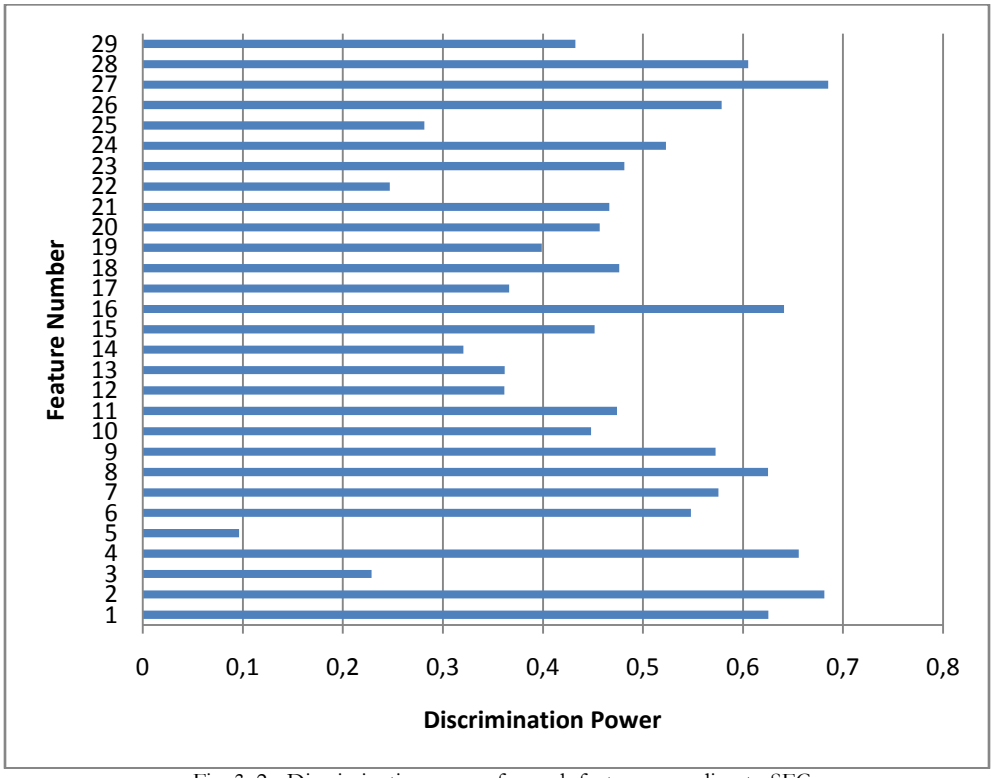


Fig. 3. 2 - Discrimination power for each feature according to SFC

Although, the feature ranking is considered to be a non-compulsory process, it gives more information about the value of each feature according to the problem at hand. Clearly, the feature ranking process represents a coarse filtration of the features. For instance in Fig 3.1, Fig. 3.2, one can see that some features have a low discriminations power, so we may discard these features afterwards.

From Fig. 3.3, one can see very clear that CBR is similar (sometimes superior) to SFC. For the special case of AD-LBD, the CBR gives a performance which is higher than the one of SFC.

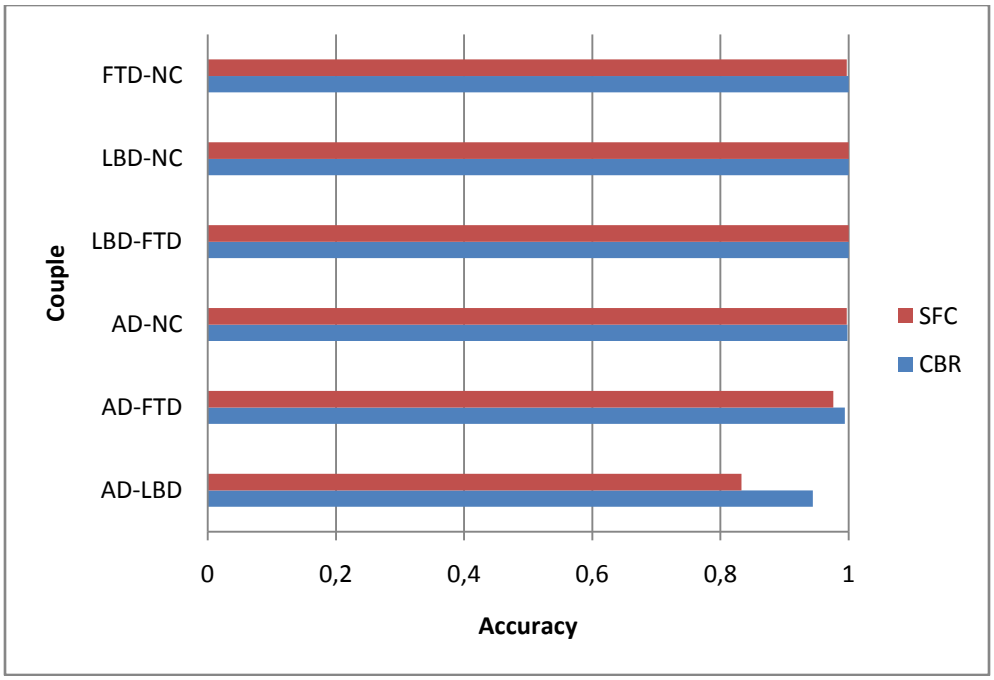


Fig. 3. 3 - Comparison of the classifier performance for CBR-based classification and SFC-based classification respectively

In addition, the CBR take just some seconds to generate the ranked list of features while the SFC takes more time as it uses the classifier itself for testing the influence of every single feature. Thus CBR is a good choice, in particular when considering a large number of features.

After ranking the features, the task is not completed yet; the influence of class imbalance on the classifier training process is studied.

C h a p t e r 4

THE INFLUENCE OF CLASS IMBALANCE ON TRAINING

Summary

C h a p t e r 4

THE INFLUENCE OF CLASS IMBALANCE ON TRAINING

4.1 Introduction

Most of today's research on classification training has focused on the training algorithm itself. However, it is intuitively evident that imbalance in the class distribution might lead to significant biases in the final classification. Furthermore, some studies have showed that the class distribution has a significant effect on the training process [11, 12].

Various research studies try to reduce the large data sets, in size, before training the classifier in order to decrease the processing time. Hence, one can feel that it's important to study the influence of class distribution on the classifier learning. Although, it is not the primary aim in problem at hand (as there's no large number of data sets), it was also important to study the class distribution effect in presented work. In particular, this problem was observed in the training data for the current problem in the case of discriminating AD-LBD. Discriminating LBD from AD is not an easy task and this is because of:

(i) The available number of AD cases (44) is more larger than the available number of the LBD cases (9)
(ii) AD and LBD have similar disease patterns

According to the problem at hand arises an important question: What is the optimal number of data sets for training? or in another way, what is the optimal number of data sets to improve the classifier learning? To answer this question many solutions come to mind, e.g. (i) decrease the number of the largest data sets to the smallest number of data sets (*under-sampling*)?, or (ii) increase the number of the smallest data sets to the largest number of data sets (*over-sampling*)? A further alternative would be to simply use the "natural" distribution.

In the next section the influence of these suggested class distributions on the classifier training is studied.

4.2 Under-sampling vs. Over-sampling

Here, the most familiar sampling plans and how they affect the decision of the classifier are studied. In particular, *under-sampling* and *over-sampling*, which decrease and increase, respectively, the occurrence of one class in the training set to overcome the misclassification errors are evaluated. These two approaches are eye-catching because they reshape the training data without modifying the classification algorithm itself.

A receiver operating characteristic (ROC) analysis was performed, as usual, to compare results. Still, a question before proceeding: Which data sets to be removed (added)? Although this data removal (addition) is done randomly, one can also do this manually (according to the quality of available data) or according to any other criterion which could be considered suitable.

Tab. 4.1 and Tab. 4.2 show the different distributions used to compare *under-sampling* to *over-sampling* respectively.

Problem	AD	LBD	FTD	NC
AD-LBD	9	9	-	-
AD-FTD	13	-	13	-
AD-NC	26	-	-	26
LBD-FTD	-	9	9	-
LBD-NC	-	9	-	9
FTD-NC	-	-	13	13
AD-LBD-FTD-NC	9	9	9	9

Tab. 4. 1 - Under-sampling class distribution

Problem	AD	LBD	FTD	NC
AD-LBD	44	44	-	-
AD-FTD	44	-	44	-
AD-NC	44	-	-	44
LBD-FTD	-	13	13	-
LBD-NC	-	26	-	26
FTD-NC	-	-	26	26
AD-LBD-FTD-NC	44	44	44	44

Tab. 4. 2 - Over-sampling class distribution

4.3 Natural Distribution vs. Balanced Distribution

Learning from a *balanced* class distribution makes the classifiers, normally, arise with fewer but more accurate classification results for the minority class than for the majority class [11]. Although, it seems to be surprising that the balanced distribution results in more accurate results than the natural distribution, this surprising fact exists because the minority class often comprises a more homogenous set of entities, while the majority class often contains significantly more inter-subject variability.

Tab. 4.3 and Tab. 4.4 show the different distributions used to compare *natural distribution* to *balanced distribution* respectively.

Problem	AD	LBD	FTD	NC
AD-LBD	44	9	-	-
AD-FTD	44	-	13	-
AD-NC	44	-	-	26
LBD-FTD	-	9	13	-
LBD-NC	-	9	-	26
FTD-NC	-	-	13	26
AD-LBD-FTD-NC	44	9	13	26

Tab. 4. 3 - Natural class distribution

Problem	AD	LBD	FTD	NC
AD-LBD	9	9	-	-
AD-FTD	13	-	13	-
AD-NC	26	-	-	26
LBD-FTD	-	9	9	-
LBD-NC	-	9	-	9
FTD-NC	-	-	13	13
AD-LBD-FTD-NC	9	9	9	9

Tab. 4. 4 - Balanced class distribution

4.4 Discussion

Fig.4.1 shows that classification using *under-sampling* class distribution beats classification using *over-sampling*. On the other hand, Fig. 4.2 shows that classification using *balanced* class distribution beats classification using *natural* class distribution.

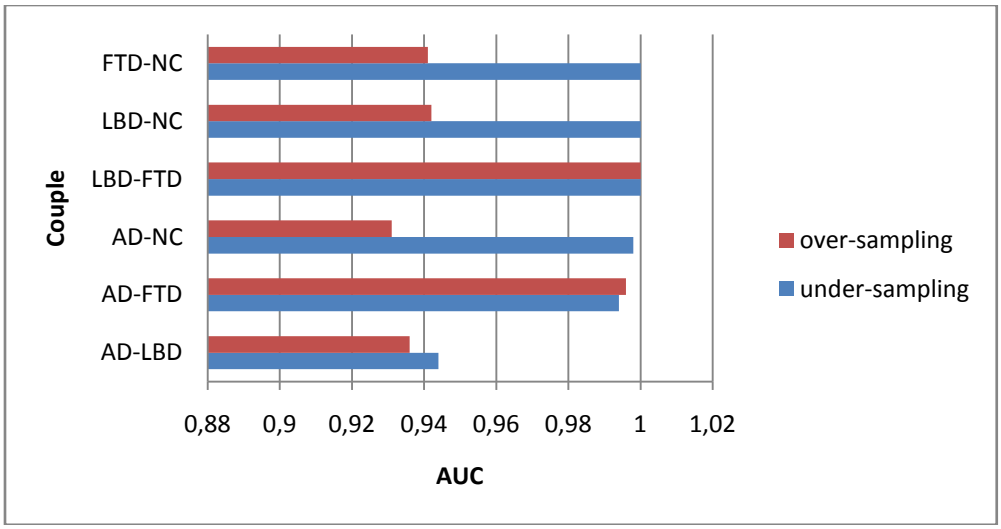


Fig. 4. 1 - AUC for all 2-class combinations using over-sampling and under-sampling class distributions

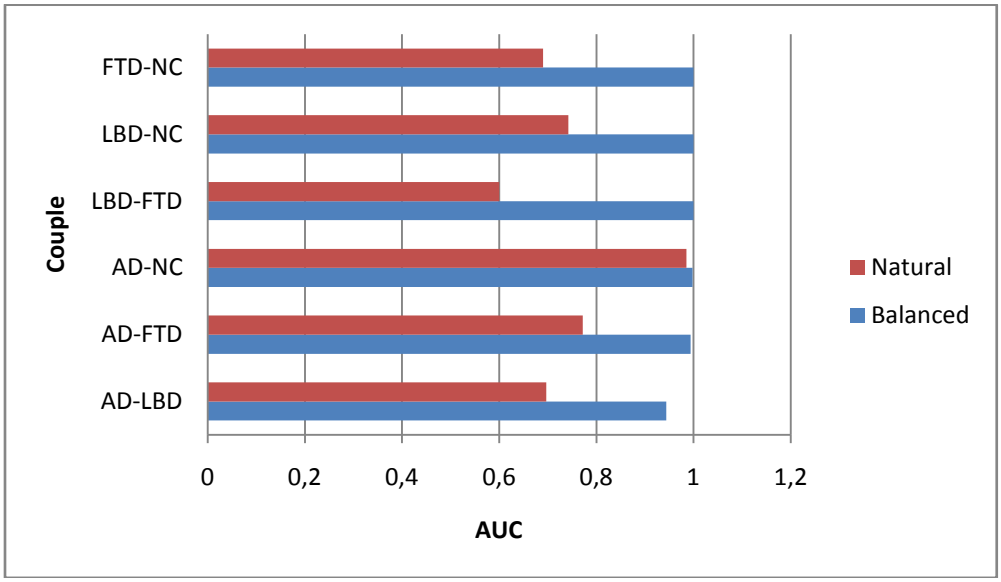


Fig. 4. 2 - AUC for all 2-class combinations using natural and balanced class distributions

The result shows that training the classifier with a natural class distribution cause the classifier to likely have higher error rate for the minority class. The main reason for this behavior of the classifier is that the minority class has fewer training samples than the majority one and that makes the classifier biased to predict the majority class most of the time.

It is clear now that the class distribution influences the classifier training. This is because the training with a balanced class distribution increases the prediction of the minority class. Despite that increase in the minority class is slight; this improves the whole prediction accuracy.

In the next chapter some of the feature selection methods and how the feature ranking process can help in optimizing the feature selection process are discussed.

C h a p t e r 5

FEATURE SELECTION

Summary

C h a p t e r 5

FEATURE SELECTION

5.1 Introduction

In theory, more features should give more discriminating power. In spite of that hypothesis, in reality, with an inadequate number of training data, the discriminating power is higher as excessive features are omitted. Excessive features also notably slow the training process, though this has no effect on discrimination power. Furthermore, excessive features may also cause the classifier to over-fit the training data as irrelevant features may mislead the classifier.

As features are already generated (extracted), the feature selection approach does not attempt to generate new features, alternatively it tries to select $d < D$ features by getting rid of irrelevant and redundant features and maintaining the relevant ones. (See Fig. 5.1) The feature selection can have many influences on the automatic classification problem:

(i) Giving insight into the nature of the classification problem
(ii) Improving of the classification accuracy
(iii) Simple and potentially faster classifier can be built easily

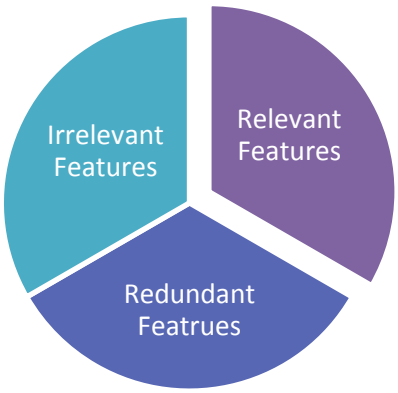


Fig. 5. 1 - A view of feature relevance

To make it clearer, let X_D be the set which includes all the D original features, and χ_d the set of all possible subsets of size d (where d is the desired number of features). Furthermore, let X be a subset of X_D with a number of features which is equal to d, and $J(X)$ be an evaluation function which quantifies the quality of the subset X. By definition, a higher value of J means a better feature subset. Now, the feature selection problem is to be seen as a searching problem that searches $\tilde{X}_d \subset X_D$ of d features to satisfy:

$$J(\tilde{X}_d) = \max_{X \in \chi_d} J(X) \qquad \text{(Eq. 5.1)}$$

As feature selection plays an important role in the classification process, diverse methods already exist and are used for feature selection. Feature selection algorithms can be categorized with respect to search strategies or selection criteria [18] (see Fig. 5.2). Diverse search strategies have been studied to select optimal features for evaluation, in order to have a balanced trade-off of accurate result and efficient computation.

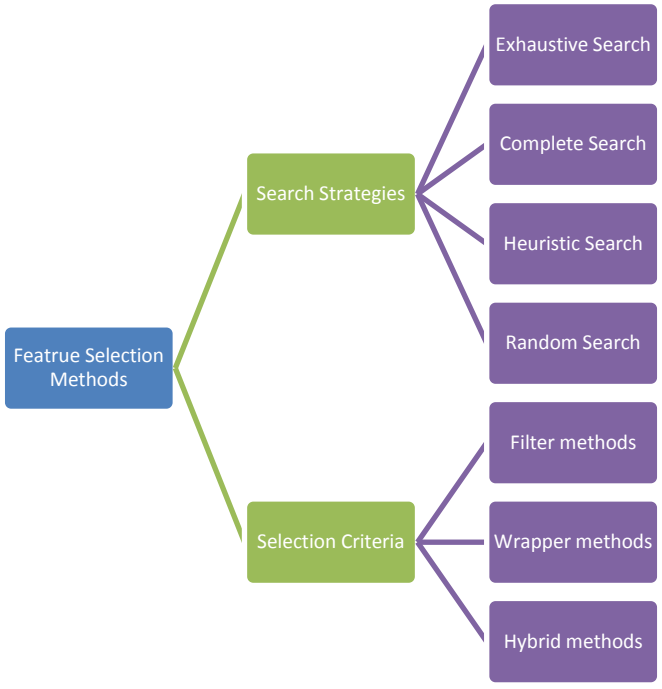


Fig. 5. 2 - Categorization of feature selection methods

Feature Selection Search Strategies

The *Exhaustive Search* approach entails investigating all possible subsets of size $d < D$ of the original set X_D and selecting the subset with the highest J value. It is obvious that this approach guarantees an optimal solution, but it is still a computationally intensive strategy [18].

Complete Search is an alternative approach to the exhaustive one and it's also known as *Branch & Bound* algorithm (B&B) [18]. It's based on the *monotonicity property* of the feature selection criterion: having two subsets of the original set X_D, A and B, if $A \subset B$, then $J(A) < J(B)$. Therefore, according to this property, many feature subsets can be ignored. Although, this search approach guarantees the selection of an optimal subset of size d, the computation time has exponential direct proportion with data dimensions.

Heuristic Search is a non-exhaustive selection approach. Diverse heuristic methods have been proposed, e.g. Sequential Forward and Backward Selection, Plus-*l*-Take Away-*r*, Genetic algorithms, and the Floating and Oscillating Search. The computation time of these methods has quadratic (or less) direct proportion with data dimensions.

Random Search, as implied by the name, a random search strategy. It begins with a subset of features in which all the features in this seed subset are randomly chosen. Then, some features are added or removed, also randomly, till the termination at a specific number of iterations. The computation time can have a linear relationship with the iterations number.

Some heuristic search algorithms try to make use of the random search principle as a catalyst to improve the heuristic search in some ways. For instance, one can use the random search principle to choose a random subset of features to be the seed for the heuristic search algorithm.

Feature Selection Criterion Functions

Filter approach is based on the separation of the performance evaluation function from the classifier (See Fig. 5.3). More clearly, the performance evaluation function is calculated directly from the training data sets, i.e. *correlation*.

Wrapper approach uses the performance of the classifier directly as an evaluation function (See Fig. 5.3). The intent is to find an optimal subset of features that suit well to the classifier. The main advantage of using the wrapper approach is that the classifier error rate is used directly to evaluate the

performance. This means that the evaluation error space is the same as the classifier error space.

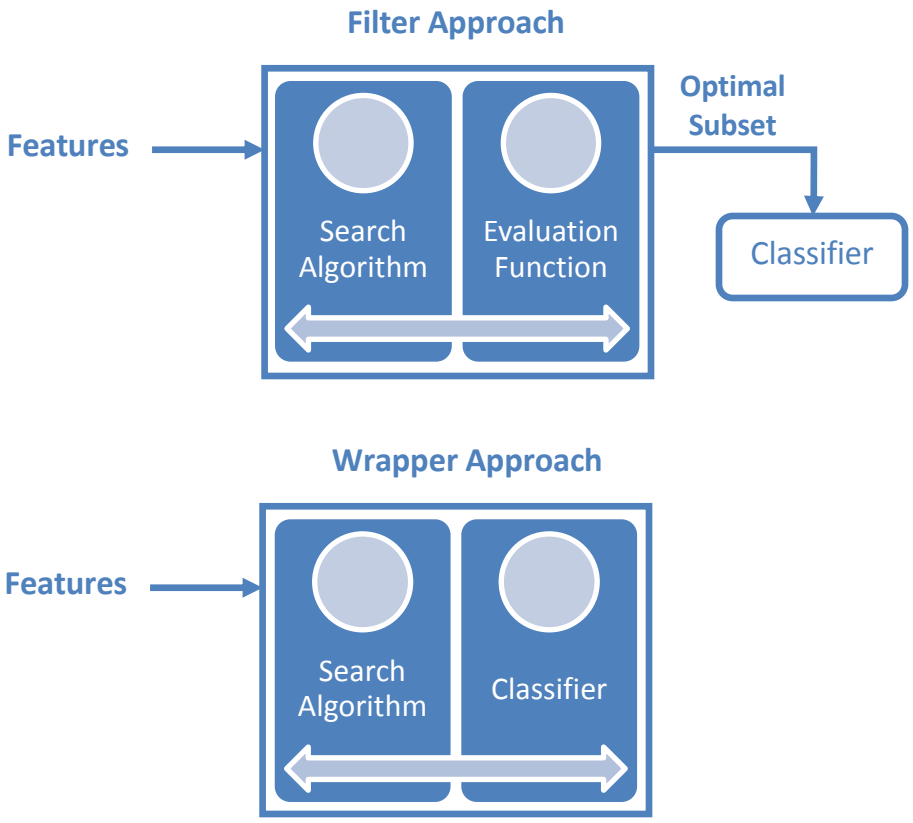


Fig. 5. 3 - Common criterion approaches for feature selection

On the other hand, although the wrapper method achieves better performance (than the filter method), it is computationally more intensive because it uses the classifier itself to perform the evaluation. Another drawback is that the wrapper approach gives a subset which is mainly optimized for the current classifier and maybe it is not the best subset for another classifier.

Hybrid approach is used to reduce the discrepancy between filter and wrapper approaches in terms of computational cost. Beside that it gathers the advantages of both approaches; it has been proposed to work with high dimensional data. The computation time of the hybrid approaches is compared to that of the filter approaches [18].

In the next sections, some of the currently best-performing feature selection algorithms, and a proposed one as well are discussed:

(i) **C**orrelation-based **F**orward **S**election (**CFS**)
(ii) **S**equential **F**orward **S**election (**SFS**)
(iii) **S**equential **F**loating **F**orward **S**election (**SFFS**)
(iv) **G**oodness-based **S**equential **F**loating **F**orward **S**election (***g*SFFS**)

Before proceeding to explain these methods in detail, Tab. 5.1 provides an overview of each method, and the categories in which it belongs.

Algorithm	Search Strategy	Criterion Approach
CFS	Heuristic	Filter
SFS	Heuristic	Wrapper
SFFS	Heuristic	Wrapper
gSFFS	Heuristic	Hybrid

Tab. 5. 1 - Four types of feature selection algorithms

5.2 Correlation-based Feature Selection (CFS)

CFS uses a search strategy and a criterion function to evaluate the merit of the feature subsets, like any other feature selection algorithm. CFS measures the goodness G of feature subsets taking into account the inter-correlation between individual features as well as the usefulness of these features for predicting the class label. We can state the hypothesis on which the heuristic is based as:

"Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other" [19]

The above heuristic can be formalized according to:

$$G_i = \frac{D\bar{r}_{oi}}{\sqrt{D + D(D-1)\bar{r}_{ii}}} \qquad \text{(Eq. 5.2)}$$

where G_i is the goodness of feature i, D is the total number of features, $\bar{r}_{oi}$ is the mean value of the outer-correlation (correlation with the class) for feature i, $\bar{r}_{ii}$ is the average feature inter-correlation (mutual correlation between feature i and all other features).

Eq. 5.2 was used by *Hall* [19] which is borrowed from test theory according to *Ghiselli* [20], where it is supposed to be used to measure the reliability of a test consisting of summed items from the reliability of the individual items. For instance, a more truthful indication of a person's occupational success can be obtained from a composite of a number of tests measuring a broad range of traits (leadership, academic ability etc), more willingly than any single individual test that measures a limited scope of trait.

Hall reused this equation, but for the purpose of feature subset selection. A closer look at goodness equation reveals that the equation is, in fact, a standardized Pearson's correlation. The goodness equation filters out irrelevant features, as they are poorly correlated with the class (numerator). On the other hand, redundant features shall be disregarded because they are highly correlated with one or more other features (denominator).

CFS Algorithm

Input:

$$Y = \{y_i \mid j = 1, \dots, D\}$$

Output:

$$X_k = \{x_j \mid j = 1, \dots, k,\ x_j \in Y\}, k = 0,1, \dots, D$$

Initialization:

$$X_0 := 0;\ \ k := 0$$

Termination:

$$Stop\ when\ k\ equals\ the\ number\ of\ features\ required$$

Step 1 *(Goodness Calculation)*

$$G_x = \frac{D\bar{r}_{xo}}{\sqrt{D + D(D-1)\bar{r}_{xi}}}$$

Step 2 *(Inclusion)*

$$x^+ := \arg \max_{x \in Y - X_k} |G_x|$$

$$X_{k+1} := X_k + x^+;\ \ k := k + 1$$

Fig. 5. 4 - CFS Algorithm

Fig. 5.4 shows the CFS algorithm which consists of two main steps:

(i) Goodness calculation G, which consists of two inner steps:
   - average outer-correlation calculation $\bar{r}_{oi}$
   - average inter-correlation calculation $\bar{r}_{ii}$

(ii) Inclusion, which includes the number of required features into your empty subset according to their goodness value.

5.3 Sequential Forward Selection (SFS)

The Sequential Forward Selection (SFS) and its backward equivalent (SBS) are well-known suboptimal methods to find a sequence of nested subsets of features in a straightforward manner, i.e. by adding (subtracting) the locally best (worst) feature in the set [21]. SFS was first used by *Whitney* [22] to determine the best subset of measurements out of a set of D total measurements. SFS gained further prominence when *Pudil* [23] adapted it in his SFFS algorithm.

In details, SFS starts the search with an empty set. In each step, every feature is considered for selection unless it has been already selected, and the estimated classification accuracy is recorded. At the end of the step, the feature to be included in the set is the feature whose inclusion resulted in the best accuracy. Afterwards, a new step is started, remaining features are considered, and so on. This algorithm terminates when a defined number of features is included.

The main drawback of the SFS is the nesting property, which means that included feature cannot be excluded later, even though such a strategy might boost the accuracy [24].

SFS Algorithm

Input:

$$Y = \{y_i \mid j = 1, \dots, D\}$$

Output:

$$X_k = \{x_j \mid j = 1, \dots, k,\ x_j \in Y\}, k = 0,1, \dots, D$$

Initialization:

$$X_0 := 0$$

Repeat

$$x^+ := \arg \max_{x \in Y - X_k} J(X_k + x)$$

$$X_{k+1} := X_k + x^+;\ k := k + 1$$

Until *no improvement in J in last n steps* ***or*** $k = D$

Fig. 5. 5 - SFS Algorithm

5.4 Sequential Floating Forward Selection (SFFS)

The Sequential Floating Forward Selection (SFFS) and its backward equivalent (SFBS) were first introduced by *Pudil* [23]. In many comparisons [21, 23, 25], the SFFS has proven to be superior to the SFS algorithm.

The main idea behind the SFFS is to face up the problem of nesting that we had already with the SFS. This is done in the SFFS algorithm by starting an exclusion process (excluding features) after the inclusion of one feature. The exclusion goes on as long as there's a better feature subset which can be found. This exclusion stops when there's no better feature subset, of the equivalent size, can be found and the algorithm goes back to the first step and includes the currently excluded feature, which is again followed by exclusion process and so on.

SFFS Algorithm

Input:

$Y = \{y_i \mid j = 1, \dots, D\}$

Output:

$X_k = \{x_j \mid j = 1, \dots, k,\ x_j \in Y\}, k = 0,1, \dots, D$

Initialization:

$X_0 := 0;\ k := 0$

Termination:

$Stop\ when\ k\ equals\ the\ number\ of\ features\ required$

Step 1 *(Inclusion)*

$x^+ := \arg \max_{x \in Y - X_k} J(X_k + x)$

$X_{k+1} := X_k + x^+;\ k := k + 1$

Step 2 *(Conditional Exclusion)*

$x^- := \arg \max_{x \in X_k} J(X_k - x)$

$if\ \ J(X_k - \{x^-\}) > J(X_{k-1})\ \ then$

$X_{k-1} := X_k - x^-;\ k := k - 1$

$go\ to\ \textbf{step 2}$

$else$

$go\ to\ \textbf{step 1}$

Fig. 5. 6 - SFFS Algorithm

Fig 5.7 shows a simplified flowchart for the SFFS algorithm according to [26].

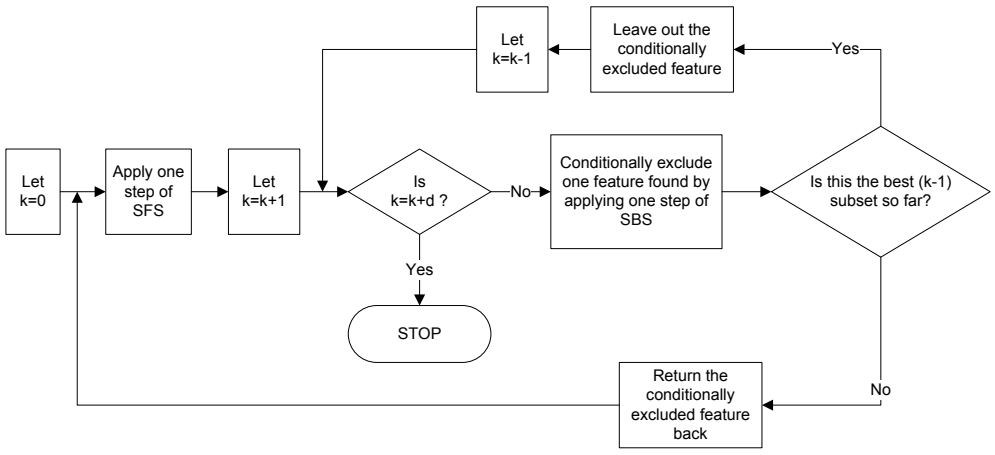


Fig. 5. 7 - Simplified flowchart of the SFFS algorithm

Although, the SFFS is more complex and the search takes more time, one can obtain better feature subsets. The original SFFS algorithm has two flaws:

(i) Repetition
(ii) Infinite-loop

The *Repetition* problem arises from repeating to estimate the classification accuracy of a subset that was already estimated before. This is not a harmful thing, but it consumes more execution time especially when there are a lot of features there. This repetition problem can be solved potentially in a straight forward manner by reconstructing a look-up table which has the records for all previous estimations. Hence, the algorithm will only estimate the accuracy for a given subset if it was not estimated previously.

The *Infinite-loop* problem arises when you have an optimal subset in the first step (inclusion) by adding a specific feature and then in the second step (exclusion) you find that the by removing another specific feature the accuracy improves. At this moment, the algorithm will repeat this process and hence we get an infinite-loop. This problem can be solved by many ways, for instance, by stopping the algorithm when there's no better accuracy achieved in the last n steps.

5.5 Goodness-based Sequential Floating Forward Selection (gSFFS)

A new feature selection algorithm is proposed here, Goodness-based Sequential Floating Forward Selection (*g*SFFS), which is a hybrid algorithm based on CFS and SBS (See sections 5.2, 5.3). The basic idea of the *g*SFFS consists of using CFS to find good feature subsets (features highly correlated with class, uncorrelated with each other) and then using SBS to try to minimize the best subset without degrading the best accuracy.

gSFFS Algorithm

Input:

$Y = \{y_i \mid j = 1, \dots, D\}$

Output:

$X_k = \{x_j \mid j = 1, \dots, k,\ x_j \in Y\}, k = 0,1, \dots, D$

Initialization:

$X_0 := 0;\ \ k := 0, B := 0\ \ where\ B\ is\ the\ best\ accuracy\ ever$

Termination:

$Stop\ when\ k\ equals\ the\ number\ of\ features\ required$

Step 1 *(Goodness Calculation)*

$$G_x = \frac{D\bar{r}_{xo}}{\sqrt{D + D(D-1)\bar{r}_{xi}}}$$

Step 2 *(Conditional Inclusion)*

$x^+ := \arg \max_{x \in Y - X_k} |G_x|$

$if\ \ J(X_k + \{x^+\}) > B\ \ then$

$X_{k+1} := X_k + x^+;\ \ k := k + 1;\ B := J(X_k + \{x^+\})$

$go\ to\ \mathbf{step\ 2}$

$else$

$Y := Y - x^+$

$go\ to\ \mathbf{step\ 1}$

Step 3 *(Conditional Exclusion)*

$x^- := \arg \max_{x \in X_k} J(X_k - x)$

$if\ \ J(X_k - \{x^-\}) \geq B\ \ then$

$X_{k-1} := X_k - x^-;\ \ k := k - 1;\ B := J(X_k - \{x^-\})$

$go\ to\ \mathbf{step\ 3}$

Fig. 5. 8 - gSFFS Algorithm

The first step (*Goodness calculation*), is done by calculating the *Goodness* (See section 5.2) for all the features. In the second step (*Conditional Inclusion*), we add to the optimal subset, which starts empty, the feature which has the highest goodness value, if and only if, this results in increasing the best accuracy which is set to `zero` at the start. In the last step (*Conditional Exclusion*), we try to minimize the optimal subset by applying SBS on a condition that the best accuracy is not degraded. Fig. 5.9 shows a simplified flowchart of the *g*SFFS algorithm.

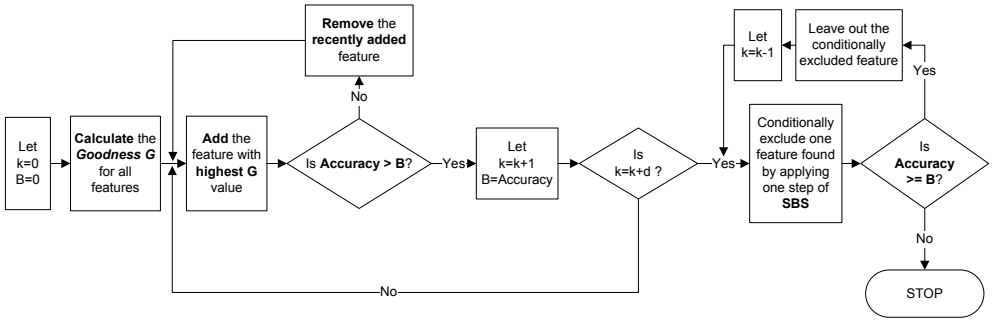


Fig. 5. 9 - Simplified flowchart of the gSFFS algorithm

Even though the *g*SFFS algorithm gives results which are comparable to the SFFS one, the *g*SFFS is superior to the SFFS as it takes significantly less time to find the best subset. This is clear as the SFFS sends each available feature to the classifier to check for the accuracy, but the *g*SFFS avoids this pitfall by calculating the goodness of the features, which is significantly less computing expensive. On the other hand, the SFFS performs the SBS in each iteration which also consumes much time, but in *g*SFFS, the SBS is applied only one time after finding the best feature subset ever.

In the next section, the results of using these different selection algorithms to find the best one suitable for the discrimination of dementia diseases, according to the current setup are demonstrated.

5.6 Feature Selection Algorithms Evaluation

The feature selection algorithms are evaluated in two different ways. First, algorithms are evaluated through the accuracy obtained with balanced training data and optimal feature subset (LOOCV). Second, algorithms are evaluated by

testing for all data sets, as some data were already removed in the balancing, and selected feature subset (Testing).

Fig. 5.10 shows the AUC for all 2-class problems using the described feature selection methods. One can see that wrapper (SFS and SFFS) and hybrid (*g*SFFS) methods perform better than the filter method (CFS). Moreover, wrapper and hybrid methods give comparable performance through the LOOCV.

Fig 5.11 shows the AUC for testing the classifier for all available cases using the features selected in the LOOCV phase. It is clear that the wrapper and hybrid methods still give better performance. It is astonishing that the classical SFS method is superior to all other methods for the special case of discriminating LBD from AD.

It is not so easy to prefer one method to others as it depends on the application. For instance, for some applications SFS could perform better, but in other applications SFFS or *g*SFFS could be the best choice. In addition, it is not only a matter of accuracy, but also time plays an important role in this feature selection process.

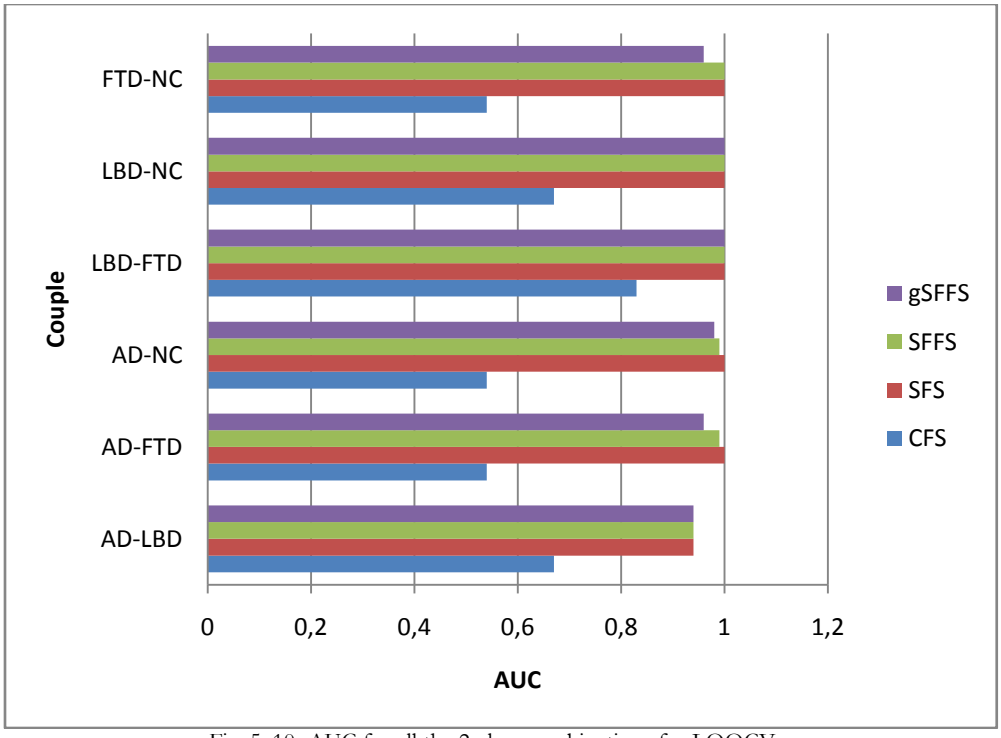


Fig. 5. 10 -AUC for all the 2-class combinations for LOOCV

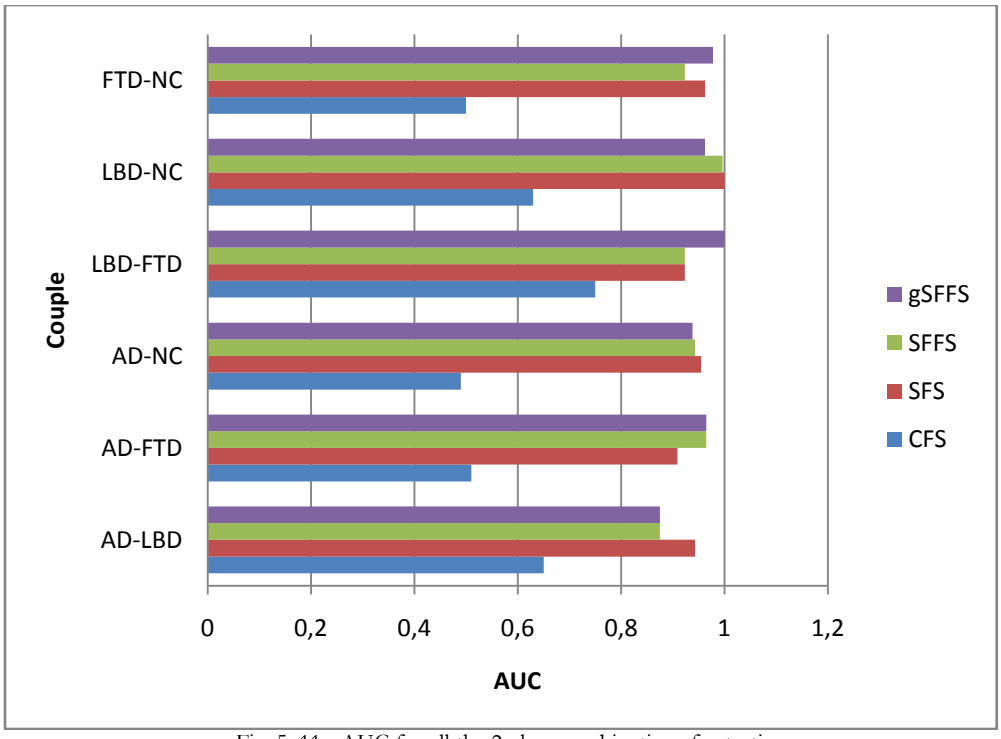


Fig. 5. 11 - AUC for all the 2-class combinations for testing

Fig. 5.12 shows the search time[5] that was needed for finding the optimal subset of feature by each algorithm. Of course, the CFS required minimum search time as any other filter method. Then, *g*SFFS takes a little bit more time between 0.5 to 4.5 minutes to find the best subset of features. After that come SFS which takes more time and at last comes SFFS which takes long time between 0.75 to 4.5 hours.

Consider also that optimal number of features found by the algorithm. Fig. 5.13 shows the optimal number of features found by each algorithm. SFFS and *g*SFFS give an optimal number of features between 1-3 features. Comparing these numbers to the ones found by CFS and SFS (1-12 features), SFFS and *g*SFFS are giving an optimal number of features with an excellent accuracy.

Looking more deeply at SFFS and *g*SFFS, *g*SFFS seems to be superior, let us explain why. *g*SFFS found the same number of features for all 2-class problems (although the subsets were not the same in all cases). In addition, *g*SFFS

[5] The demonstrated results are obtained using x86 *Intel*® machine with dual core *Xeon*™ CPU (2.4GHz) and 4GB of RAM under *Linux* OS. The programming was made in *C++* (programming) and *Python* (scripting). *Boost C++ libraries* were used for efficient mathematical calculations.

gives a comparable AUC to SFFS with an average difference of ±0.0155. Finally, it was observed already that *g*SFFS takes an average of 1.4 minutes, while the SFFS takes an average of 158.5 minutes.

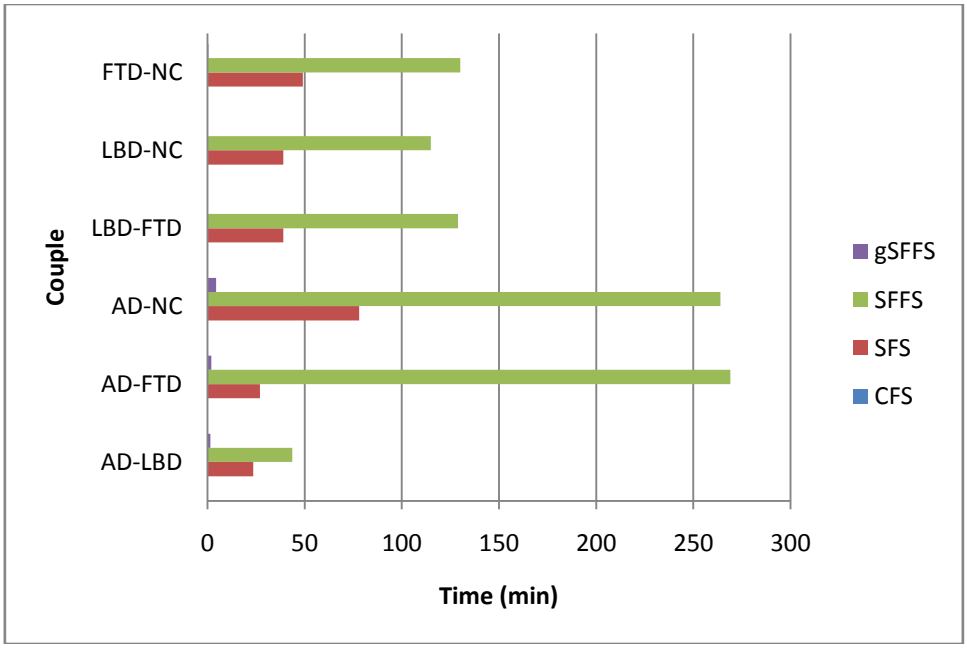


Fig. 5. 12 - Search time for all 2-class problems

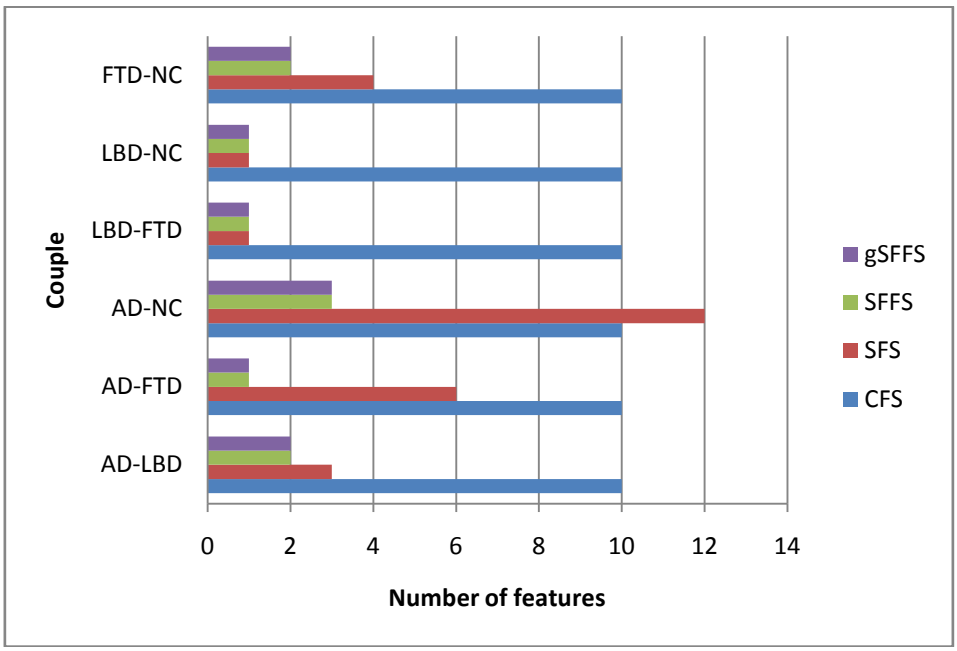


Fig. 5. 13 - Optimal number of features found by each algorithm for all 2-class problems

In conclusion, it is clear that the feature selection problem is a compromise between accuracy and computation time. For some applications, computation time will be of primary importance and for others; accuracy will be the most important regardless of the search time.

The results presented illustrate that hybrid methods offer significant advantages, while not compromising on accuracy. Hybrid methods take less time while providing good accuracy. If hybrid methods are not available, then wrapper methods would be a good choice.

In the next chapter, the results of all approaches which applied in this work, from feature extraction till the effect of imbalanced class distribution on classifier training, are concluded.

C h a p t e r 6

CONCLUSION & FUTURE WORK

Summary

6.1 Conclusion

In conclusion, it is clear now that the methods used can improve the classification of FDG-PET in patients with dementia. These methods can be integrated in one methodology that is capable of significantly improve the classification problem at hand. This methodology is to be called *BRST* which stands for ***B***alance-***R***ank-***S***elect-***T***rain.

Fig. 6. 1. BRST methodology

Using a balancing training data prevents bias during classifier training. Feature ranking has not a direct effect on improving the classification task, though it can reduce the search time in the feature selection step. Feature selection hits two birds with one stone. First, it reduces the future training process by selecting an optimal subset of features. Second, it improves the accuracy by eliminating irrelevant and redundant features. Finally, future classifier training will result in obtaining higher classification accuracy.

Balancing training data	prevents bias during training
Ranking of features	reduces computation time
Selection of optimal features	increases discrimination accuracy
Training with balanced data and optimal features	optimize future classification results

Tab. 6. 1. The influence of the BRST methodology on the classification process

Although that one is free to choose the appropriate feature selection algorithm with this methodology, the proposed feature selection algorithm (*g*SFFS) was superior to traditional feature selection algorithms (CFS and SFS), and it gave comparable accuracy to one of the best-performing algorithms (SFFS) but even in some cases it was superior. Anyway, the feature selection is a compromise between computation time and obtained accuracy, and depending on the application, one can choose the appropriate selection algorithm.

Finally, the proposed methodology (*BRST*) is straightforward method and easy to implement. It helps preventing bias, reducing computation time, and increasing the classification accuracy.

6.2 Future Work

Although the proposed methodology managed to improve the accuracy for all pair-wise (2-class) classification problems, it failed to improve the accuracy for multi-class (4-class) classification problem. Multi-class classification is still an open question and there are a lot of current researches in this area.

BIBLIOGRAPHY

[1]. *What is dementia?* **Alzheimer's.Australia.** 2005, Vol. Help Sheet 1.1.

[2]. *Dementia Facts & Statistics.* **Alzheimer's.Australia.** 2008, Vol. Alzheimer's Australia Statistics.

[3]. *Global prevalence of dementia: a Delphi consensus study.* **Ferri, Cleusa P, Prince, Martin and Mary, Laura Fratiglioni.** 2005, The Lancet, Vol. 366.

[4]. *Alzheimer's disease.* **Alzheimer's.Australia.** 2005, Vol. Help Sheet 1.12.

[5]. *Dementia with Lewy bodies.* **Alzheimer's.Australia.** 2005, Vol. Help Sheet 1.15.

[6]. *Fronto temporal lobar degeneration (FTLD).* **Alzheimer's.Australia.** 2005, Vol. Help Sheet 1.14.

[7]. *Early Diagnosis of Dementia.* **Alzheimer's.Australia.** 2007, Vol. Paper 10.

[8]. *Neuroimaging and Early Diagnosis of Alzheimer Disease: A Look to the Future.* **Petrella, Jeffrey R., Coleman, R. Edward and Doraiswamy, P. Murali.** 2003, Radiology, Vol. 226, pp. 315-336.

[9]. *Automatic classification of FDG PET brain scans for dementia with partial least squares.* **Wenzel, F. and Young, S.** Philips Research Europe - Hamburg. 2007.

[10]. *B-Spline Registration of 3D Images with Levenberg-Marquardt Optimization.* **Kabus, S, et al.** 2004, SPIE Medical Imaging.

[11]. *The Effect of Class Distribution on Classifier Learning.* **Weiss, Gary M. and Provost, Foster.** Department of Computer Science, Rutgers University. 2001.

[12]. *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling.* **Drummond, Chris and Holte, Robert C.** 2003, Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC.

[13]. *Towards multimodal atlases of the human brain.* **Toga, Arthur W., et al.** 2006, Nature Reviews Neuroscience, Vol. 7, pp. 952-966.

[14]. *Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain.* **Tzourio-Mazoyer, N., et al.** 2002, Neuroimage, Vol. 15, pp. 273-289.

[15]. *3D statistical neuroanatomical models from 305 MRI volumes.* **Evans, A. C., et al.** 1993, Proc. IEEE-Nuclear Science Symposium and Medical Imaging Conference, pp. 1813-1817.

[16]. *Digitaler VOI-Atlas für SPECT und PET des Gehirns in der klinischen Patientenversorgung.* **Wilke, F., et al.** 2008, der Deutschen Gesellschaft für Nuklearmedizin 46.

[17]. *An Introduction to Variable and Feature Selection.* **Guyon, Isabelle and Elisseeff, Andre.** 2003, Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

[18]. *Notes on the evolution of fetaure selection methodology.* **Somol, Petr, Novovicova, Jana and Pudil, Pavel.** 2007, Kybernetika - The Journal of the Czech Society for Cybernetics and Information Sciences, Vol. 43, pp. 713-730.

[19]. *Practical feature subset selection for Machine Learning.* **Hall, M. and Smith, L.** Proceedings of the Australian Computer Science Conference (University of Western Australia), 1996.

[20]. *Theory of Psychological Measurements.* **Ghiselli, E. E.** McGraw-Hill, 1964.

[21]. *Comparative study of techniques for large-scale feature selection.* **Ferri, F., et al.** 1994.

[22]. *A Direct Method of Nonparametric Measurement Selection.* **Whitney, A. W.** 1971, IEEE Trans. on Computers, Vols. C-20, pp. 1100-1103.

[23]. *Floating Search Methods In Feature-Selection with Nonmonotonic Criterion Functions.* **Pudil, P., Novovicova, J. and Kittler, J.** 1994, PRL, Vol. 15, pp. 1119-1125.

[24]. *Overfitting in Making Comparisons Between Variable Selection Methods.* **Reunanen, Juha.** 2003, Journal of Machine Learning Research, Vol. 3, pp. 1371-1382.

[25]. *Featrue Selection: Evaluation, Application, and Small Sample Performance.* **Jain, Anil and Zongker, Douglas.** 1997, IEEE Trans. Pattern Anal. Machine Intell, Vol. 19, pp. 153-158.

[26]. *Adaptive floating search methods in feature selection.* **Somol, P., et al.** 1999, Pattern recognition Letters, Vol. 20, pp. 1157-1163.

[27]. *An introduction to ROC analysis.* **Fawcett, Tom.** 2006, Pattern recognition Letters, Vol. 27, pp. 861-874.