Department of Mathematics

Master's Thesis

Leiden University

Dissimilarity based learning

Niels Jongs

1st Supervisor: Prof. dr. Mark de Rooij
2nd Supervisor: Dr. Tim van Erven
3rd Supervisor: Prof. dr. Nic van der Wee

2 December 2015

Universiteit
Leiden
The Netherlands

Leiden University Medical Center

Acknowledgements

First of all, I would like to convey my gratitude to my first supervisor Professor Mark de Rooij for his guidance, generous input and sharing of knowledge, valuable comments and encouragement until the end of my thesis. Thank you for introducing me into the fascinating topic of dissimilarities and how they are of additional value in statistics. I would also like to express my sincere gratitude to my second supervisor Dr. Tim van Erven for his generous sharing of knowledge about scientific writing and the open conversations about the topic of my thesis. Additionally I would like to express my gratitude to Professor Nic van der Wee for his sharing knowledge about psychiatric disorders. Simply said, one could not have wished for friendlier or more open supervisors.

Special thanks are given to the Netherlands Study of Depression and Anxiety consortium for generous sharing of data.

The most important are my loved ones, I owe more than a simple thanks to my family members which includes my beloved parents for their financial support and encouragement throughout my study. Without their support, it was far beyond possible for me to graduate and to explore my talents.

Abstract

In this study the performance of feature-based dissimilarity space(FDS) classification is evaluated by comparing it to conventional classification techniques. In FDS classification a classifier is trained by using a dissimilarity space instead of a feature vector space. Since FDS classification is applied in a wide range of classifiers a new and model independent dissimilarity feature selection method is presented and tested. The fundamentals of this newly proposed selection method are given by the compactness hypothesis(Arkadev and Braverman, 1966). The performance of this newly proposed dissimilarity feature selection technique is evaluated by a Monto-Carlo simulation experiment and a bootstrap study.

The performance of FDS classification is evaluated by comparing it to the performance of conventional classification techniques. The performance of FDS classification is estimated by using a bootstrap procedure. The results indicate that FDS classification is beneficial in combination with a linear classifier and a complex classification task. Due to the combination of a linear classifier and FDS classification a linear decision boundary is fitted in a dissimilarity space. This decision boundary becomes non-linear in the original feature vector space.

Contents

Acknowledgments i

Abstract ii

1 Introduction 1

1.1 Aim of this thesis . . . 2

1.2 Structure . . . 3

2 Formalising Dissimilarity-Based Classification 4

2.1 Concept of Compactness in Dissimilarities . . . 4

2.2 Representation Set . . . 9

2.2.1 Prototype optimisation . . . 10

2.3 Dissimilarity Feature Selection . . . 10

2.4 Feature-Based Dissimilarity Space Classification . . . 12

2.5 Monte-Carlo Simulation Experiment . . . 12

2.5.1 Results . . . 15

2.5.2 Conclusion . . . 17

2.6 Support Vector Machine - Bootstrap Study . . . 19

2.6.1 Results . . . 22

2.6.2 Conclusion . . . 24

3 Data 26

3.1 Features . . . 27

3.2 Formulation Dissimilarity Matrix . . . 30

4 Methods 31

5 Results 36

5.1 Results Feature-Based Dissimilarity Space Classification . . . 36

5.1.1 Logistic Regression . . . 36

5.1.2 Support Vector Machine . . . 38

5.1.3 Naive Bayes Classifier . . . 40

5.1.4 Random Forest . . . 42

5.1.5 Linear / Quadratic Discriminant Analysis . . . 44

5.2 Clinical Results . . . 46

6 Conclusion **49**

6.1 Conclusion Feature-Based Dissimilarity Space Classification 49

6.2 Conclusion Clinical Implications 54

7 Discussion **57**

References **61**

1 Introduction

In statistics classification is the problem of assigning an unseen observation to a category and hereby reducing the classification error. This strongly connects classification to the discipline of machine learning. Machine learning is the study of algorithms that learn from data and has its roots in artificial intelligence but quickly became an independent discipline. In the recent years the popularity and quality of these learning algorithms has quickly increased and the number of learning algorithms has rapidly grown in the last decade. All of these learning techniques used for classifying unseen observations are based on a set of features or predictor variables, examples are logistic regression and linear/quadratic discriminant analysis.

A new classification method that is developed in the last decade uses dissimilarities to classify unseen observations and is referred to as dissimilarity-based learning. These dissimilarity-based learning methods are proposed for dissimilarity structures (Pekalska et al., 2001) and are primarily used in situations where the data consists of comparisons between objects. Dissimilarity-based learning methods are for example used in the comparison of medical images and are proven effective in classifying patients by using medical imaging data to discriminate between normal images and images with indication of disease (Arzheava et al., 2009).

More recently, feature-based dissimilarity space(FDS) classification has been proposed (Duin and Pekalska, 2006). This classification method uses dissimilarity measures in the feature vector space to represent the dissimilarities between objects and to classify unseen observations. In FDS classification the feature vector space is replaced by a dissimilarity space. A feature vector space structure is often represented by a $n \times p$ data structure with n objects and p features. This feature vector space is transformed into a dissimilarity space and is used in FDS classification to train a classification model and to classify unseen observations. This dissimilarity space defined over the original feature space consist of pairwise dissimilarities between objects and is argued to be a more natural way of representing objects (Duin and Pekalska, 2012). Goldfarb (1985) even argued that in the field of pattern recognition the use of the feature vector space should be replaced by a dissimilarity approach.

The advantage of this dissimilarity approach is the easy interpretability of the dissimilarities. If two objects are almost similar their dissimilarity measure is close to zero and thereby these objects are adjacent in their representation. This indicates that if a pairwise dissimilarity measure of two observations is zero, only if the two observations are identical, they should also belong to the same category. The use of dissimilarities in classification is supported by the argument that dissimilarity-based learning overlaps with how humans categorise objects. When humans observe an object they almost instantly determine to which class the object belongs, and most of all, humans are very accurate in this task. It is argued that this is efficient due to the fact that almost instant classification by humans is based on the perception of dissimilarities between objects and prototypes (Edelman, 1999).

FDS classification has the advantage that it can be applied in several existing classification methods, but instead of using the feature vector space, it classifies by using the dissimilarity space. Besides that, FDS classification does not require a prior specification of linearity or non-linearity. Duin et al. (2010) evaluated the performance of FDS classification by comparing it with traditional feature-based classifiers such as nearest neighbour, linear/quadratic discriminant analysis and support vector machine algorithms. They found that on a large amount of different datasets FDS classification outperforms traditional classification methods. FDS classification is also used to identify individuals with schizophrenia by using magnetic resonance imaging(MRI) data and proved to be of additional value in terms of the misclassification rate (Ulas et al., 2011). Also the early detection of dementia was done by using MRI data in combination with FDS classification (Klein et al., 2010). Both these studies are characterised by the use of complex medical data. These results indicate that FDS classification is of additional value while using complex data.

1.1 Aim of this thesis

In this thesis the performance of FDS classification will be evaluated and compared with the performance of traditional machine learning algorithms such as the Random Forest algorithm, linear/quadratic discriminant analysis(LDA/QDA), logistic regression and support vector machines. The performance of these methods will be evaluated by inspecting the misclassification rate produced while classifying unseen observations. Before evaluating FDS classification we will formalise how to apply FDS classification in practice and discuss a newly proposed dissimilarity feature selection method. Since we have no prior information about the performance of this new selection method two studies are conducted to evaluate its performance. After the formalisation of FDS classification the performance of FDS classification is evaluated by using a real world classification task. The performance of FDS classification is compared to the performance of traditional classifiers that uses the feature vector space. Estimates of the performance are obtained by using a bootstrap experiment.

The data that are used to evaluate the performance of FDS classification originates from the Netherlands Study of Depression and Anxiety(NESDA). The data mainly consists of biological parameters which are related to the presence of psychiatric disorders. These biological parameters are used to train several different classifiers, these trained classifiers are applied to identify individuals with a persistent or recurrent depression.

1.2 Structure

In section two of this document FDS classification will be formalised and a detailed description of how FDS classification is applied in a classifier is given. Section two includes two experiments in which a newly proposed dissimilarity feature selection method is evaluated. In Section three the data that is used to evaluate FDS classification are discussed and some properties of the data are given. The fourth section provides a detailed description of how the performance of FDS classification is evaluated and the collection of used classification techniques is discussed in detail. In section five the performance of FDS classification is compared to the performance of traditional classifiers. In section six and seven the results are discussed and recommendation for future research is given. In section six we conclude that FDS classification is of additional value while using a linear classifier in combination with highly complex data. In this thesis no additional value of FDS classification was observed while using a non-linear classifier or low complex data.

2 Formalising Dissimilarity-Based Classification

In this section feature-based dissimilarity space(FDS) classification will be formalised and a detailed description of how FDS classification is applied is given. First a definition of the concept of compactness will be given and its mathematical properties are discussed. In the second subsection a definition of the representation set will be given. The representation set, or the so-called prototypes or exemplars set often consists of a subset of dissimilarity features. In the third subsection we will discuss how to select an optimal subset of dissimilarity features. In the fourth subsection we will discuss the use of dissimilarities in a classification task. The final two parts of this section accommodate two studies in which the performance of a newly proposed selection method for dissimilarity features is evaluated.

2.1 Concept of Compactness in Dissimilarities

As noted in the introduction, dissimilarities are a natural way of representing the pairwise dissimilarities between observations. This is due to the belief that humans primarily use pairwise dissimilarities between concepts during the classification of objects and that the classification in terms of features comes second (Duin and Pekalska, 2012). Additional support for the use of dissimilarities in the context of a classification task is found in the compactness hypothesis (Arkadev and Braverman, 1966). The compactness hypothesis states that two almost identical objects are close in their representation in the dissimilarity space. In terms of a pairwise dissimilarity measure d, object i and j are almost identical if their pairwise dissimilarity is close to zero. Two objects are defined as almost identical if their overall difference on a set of features is small. These almost identical objects should also belong to the same category if the category labels are a function of the set of features. For two objects that are significantly different, their pairwise dissimilarity measure d is much bigger. For feature vector representations the notion of the compactness hypothesis does not hold, two completely different objects may have the same representation on feature p but they have a different outcome label and may differ on complementary features. For example, two individuals might have an identical age but do not have an identical outcome variable. These two individuals might differ on several other features. In this case the feature age violates the notion of the compactness hypothesis.

A restraint of this hypothesis is that it argues that if object i and j with representations x_i and x_j are identical, $d(x_i, x_j) = 0$, or almost similar, they should also belong to the same class. If this assumption holds the classes are perfectly separated by their dissimilarities and the misclassification rate is zero. However, in practice we often see that overlap between class labels is common. Duin (1999) proposed a compactness measure based on the compactness hypothesis to evaluate the complexity of a classification task. He argues that more complex classification problems require a larger training set and more complicated classifiers. The proposed compactness measure relaxes the assumption of no overlap between classes and provides an indication of the compactness/complexity of the classification task. Given a compactness measure c, a perfect compact observation for a classification problem occurs if $c = 1$ and the class labels are perfectly separated by a pairwise dissimilarity measure d. Duin (1999) defined the compactness measure as following *"The classes in a classification problem are compact if for an arbitrary object it is expected that its distance to an arbitrary object of the same class is smaller than its distance to an arbitrary object of another class".* The compactness measure for a set of labelled objects is estimated by using the empirical distribution and is defined as:

$$c = Pr(d(x_i, x_j) < d(x_i, x_r) \mid label(x_i) = label(x_j), label(x_i) \neq label(x_r)), \tag{2.1}$$

where $d(x_i, x_j)$ is the pairwise dissimilarity between two observations with an identical class label and $d(x_i, x_r)$ a pairwise dissimilarity between two observations with different class labels. By applying expression 2.1 to a dissimilarity matrix each pairwise dissimilarity in the dissimilarity matrix is evaluated for their contribution to the compactness of the dissimilarity matrix and is expressed in terms of probabilities. This probability is interpreted as the probability that a pairwise dissimilarity between two observations with the same label is smaller than a pairwise dissimilarity between two observations with a different label. Nevertheless, calculating a compactness measure for each individual pairwise dissimilarity is hardly informative. However, a compactness measure for each dissimilarity feature(column in dissimilarity matrix) in the dissimilarity structure is much more informative. This compactness measure for each dissimilarity feature represents an estimate of how well an observation associated with a specific dissimilarity feature differentiates between the class labels in the dissimilarity space. Let's denote the compactness measure for each dissimilarity feature as a set of $C = \{C_1, C_2, \ldots, C_n\}$ where $i = \{1, \ldots, n\}$. To formulate a set of compactness measures a set of $n^1 = \{n_1^1, n_2^1, \ldots, n_n^1\}$ and $n^2 = \{n_1^2, n_2^2, \ldots, n_n^2\}$ is required. Each n_i^1 is defined as $n_i^1 = \sum_j^n I(y_j = y_i)$ and $n_i^2 = \sum_j^n I(y_j \neq y_i)$ where y_i denotes the class label of observation i. The compactness measure for dissimilarity feature

i is formulated by the following expression:

$$C_i = \frac{1}{n_i^1} \sum_{\mathrm{j}=1}^{n_i^1} \left(\frac{\sum_{\mathrm{r}=1}^{n_i^2} I(\delta_{ij} < \delta_{ir})}{n_i^2} \right), \tag{2.2}$$

where δ_{ij} represents a pairwise dissimilarity between observation i and observation j with an identical class label; δ_{ir} represents a pairwise dissimilarity between observation i and observation r with a different class label. The part between the parentheses in expression 2.2 is equal to expression 2.1 and calculates the compactness for each individual pairwise dissimilarity with the same class label as the observation associated with the dissimilarity feature. Expression 2.2 could be expressed as the empirical expectation of the proportion of distances towards observations with an identical class label that is smaller than the distance towards observations with a different class label:

$$Pr(\,\delta_{ij} < \delta_{ir} \,|\, label(j) = y_i \neq \, label(r)). \tag{2.3}$$

The compactness measure C_i for dissimilarity feature i estimates how well a dissimilarity feature differentiates between the class labels in the dissimilarity space. For now lets assume the dissimilarity structure is of size $n \times n$, after applying expression 2.2 n compactness measures are formulated. The average of all compactness measures provides an estimate of the compactness of the dissimilarity structure, therefore, it also gives an indication of the complexity of the classification task. Let's denote the average compactness of a dissimilarity structure as $\overline{C}$ and is defined by (Duin, 1999):

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} C_i. \tag{2.4}$$

The classes are defined as compact if $\overline{C} > .5$, $\overline{C}$ is the empirical average of c in expression 2.1. Although Duin (1999) evaluated his compactness measure on several dissimilarities sets, he did not find any $\overline{C} < .5$. He also argued that if $\overline{C} > .5$, there exist a classifier with misclassification rate lower than .5. The compactness measure as just presented requires a metric distance function d and is dependent on the distance measure. The measure of complexity also depends on the variance of features, if the variance increases the complexity/compactness measure decreases. This is illustrated in Figure 2.1, in Figure 2.1A a dataset with two features and two class labels is presented. If the variance of feature p_1 is increased the amount of overlap between the class labels is increased. As a consequence the compactness measure decreases(Figure 2.1B) and the complexity of a classification task increases.

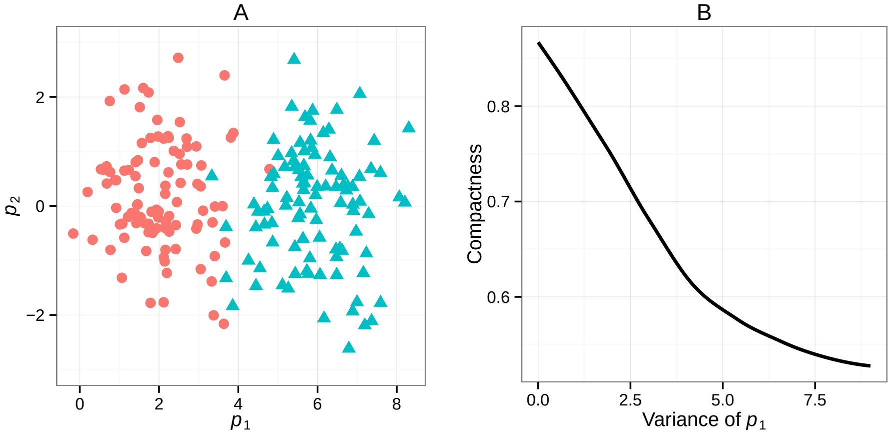


Figure 2.1: Compactness measure as a function of the variance. If the variance of feature p_1 is increased the compactness measure $\overline{C}$ decreases. As a consequence the complexity of a classification task increases.

Before we consider transforming a feature vector space into a dissimilarity space we will define the notion of metric in terms of dissimilarities. A distance measure d is defined as metric when the following conditions are satisfied:

- Non-negativity: $d(x_i, x_j) > 0$ if $x_i \neq x_j$;
- Identity of indiscernibles: $d(x_i, x_j) = 0$ if $x_i = x_j$;
- Symmetry: $d(x_i, x_j) = d(x_j, x_i)$;
- Triange inequality: $d(x_i, x_r) \leq d(x_i, x_j) + d(x_j, x_r)$;

Crucial for defining a proper dissimilarity measure are the first two conditions. The non-negativity condition states that the pairwise dissimilarity between non-identical observations is higher than zero, negative dissimilarities are considered hard to interpret. The identity of indiscernibles condition argues that a dissimilarity measure is only allowed to be zero if the two objects are identical. If this condition is violated the compactness hypothesis does no longer hold. The last two conditions are required to construct a metric dissimilarity structure. However, some argue that classification problems could also be tackled by non-metric dissimilarity structures that only satisfy the first two conditions. It is argued that dissimilarities obtained by assessing psychological constructs may not satisfy the symmetry condition (Tversky, 1977). For example, an individual might judge x towards y as more dissimilar than y towards x, as a result, the assumption $d(x, y) = d(y, x)$ is no longer satisfied. Duin and Pekalska (2010) revealed that these non-metric distances can be informative in classification problems. However, in this paper we do not deal with this problem directly since the data used is not directly observed in dissimilarities.

In this paper the dissimilarity space is defined over an original feature vector space. Suppose this feature space is represented by a dataset with n objects, $i = 1, \ldots, n$. For each object we have P representations on a set of features. $X_P = \{x_{1p},\ x_{2p}, \ldots, x_{nP}\}$. Individual measurements are denoted by lower case letters, x_{i1} represents a measurement for object i on the first feature. In this setting the data is represented in a traditional feature vector space as a $n \times P$ data-frame. In order to construct dissimilarities between objects a dissimilarity/distance function is needed that returns a scalar d_{ir} that represents the pairwise dissimilarity between object i and r on feature p:

$$d_{ir} = d(x_{ip}, x_{rp}). \tag{2.5}$$

Most commonly known dissimilarity functions are the Euclidean and the Manhattan distance function. The Euclidean distance function is defined by:

$$d_{ir} = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{rp})^2}. \tag{2.6}$$

Manhattan distance function is defined by:

$$d_{ir} = \sum_{p=1}^{P} \mid x_{ip} - x_{rp} \mid . \tag{2.7}$$

Both the Manhattan and Euclidean distance function satisfy the notion of metric dissimilarities. A less common and semi-metric distance function is the Minkowski distance function. The Minkowski distance function is defined by:

$$d_{ir} = \left(\sum_{p=1}^{P} \mid x_{ip} - x_{rp} \mid^{l} \right)^{1/l} . \tag{2.8}$$

The Minkowski distance function is metric if $l \geq 1$, and is equal to the Euclidean distance function if $l = 2$. For the $l = 1$ the Minkowski distance function is equal to the Manhattan distance function. If $l < 1$ the distance between $d(0, 0)$ and $d(1, 1)$ is $2^{1/l} > 2$ but the $d(0, 1)$ is at a distance 1 from both $d(0, 0)$ and $d(1, 1)$, this violates the notion of triange inequality and therefor the Minkowski distance is not metric with $l < 1$.

As noted earlier in this section a traditional set of features is often represented by a $n \times P$ data-matrix, after applying a dissimilarity function we obtain a dissimilarity structure of size $n \times n$. This $n \times n$ dissimilarity structure could be problematic for some traditional classifiers, therefore it is in some cases useful to reduce the size of the dissimilarity structure. The use of a representation, or so called prototype or exemplars set allow us to reduce the size of the dissimilarity structure, however there are several ways of constructing the representation set.

2.2 Representation Set

Lets assume that we have a training set T with n_T observations and given a dissimilarity measure d we have a $n_T \times n_T$ dissimilarity matrix D. Given the choice of a dissimilarity matrix D of size $n_T \times n_T$, training a classification model can be problematic for some classification methods and easily becomes computationally heavy. Therefore, a representation set $R = \{r_1, r_2, \ldots, r_h\}$ of h prototypes or exemplars is defined. Given a representation set R of size h, the size of the dissimilarity matrix D becomes $n_T \times h$ and accommodates dissimilarities between the n_T observations and the h prototypes or exemplars. Prototypes are abstract averages of the members of a class label and are non-existing. Exemplars are existing observations in the data (Ashby and Maddox, 1993) and are used to formulate the pairwise dissimilarities between n_T observations and h exemplars. In both situations the dissimilarities between observations and prototypes or exemplars is formulated, given a dissimilarity measure d a dissimilarity matrix of size $n_T \times h$ is formulated by $d(i, r)$ where $i = \{1, \ldots, n_T\}$ and $r = \{1, \ldots, h\}$. This dissimilarity matrix is used to train a classification model. The advantage of using a representation set is a decrease in computational time as noted earlier, but also the amount of noise in the data is reduced (Pekalska et al., 2006) when the prototypes or exemplars are carefully selected. There are multiple ways of defining a representation set, in the following paragraph we will discuss a method for defining a representation set based on prototypes and how to optimise these prototypes. Thereafter we will discuss how to select a subset of observations that are used as exemplars. In the case of exemplars the representation set R is defined by selecting an optimal subset of exemplars and is discussed in section 2.3. The use of a selection method is not mandatory for selecting a subset of exemplars. One option is to use all the exemplars in the training data, as a consequence the distance matrix is of size $n_T \times n_T$. An alternative option for formulating a representation set based on exemplars is to randomly select a subset of observations that are used as exemplars.

2.2.1 Prototype optimisation

As noted earlier prototypes are abstract averages of the members of a category and are non-existing. As a consequence it is possible that a prototype is 50% male and 50% female. A commonly known method for defining prototypes is the k-means clustering algorithm (Macqueen, 1967). Although several other algorithms are available, we focus on the method just mentioned. The k-means clustering algorithm is a method for finding k clusters in a unlabelled set of observations that uses the Euclidean distance. The user specifies a preferred number of k clusters and the k-means algorithms iteratively produces k clusters by minimising the within cluster sum of squares(WCSS). This k-means algorithm is applied to the feature vector data and is done before transforming the original feature vector data into a dissimilarity structure by applying a dissimilarity measure d. To formulate a set of prototypes the k-means clustering algorithm is applied to the training data for each class independently.

To use k-means clustering for defining a set of prototypes for a labeled dataset an additional step is required. If the outcome variable is a set of G labels we apply G independent k-means procedures. The following step is needed to construct prototypes:

1. Apply k-means clustering to the training data for each class independently and formulate $k = \frac{h}{G}$ clusters for each class. These clusters are the prototypes for a class label in the set G.

After these steps we have a representation set $R = \{r_1, r_2, \ldots, r_h\}$ of h prototypes. Given a dissimilarity measure d and a training set T with n_T observations we calculate the dissimilarities between the observations in the training set and the prototypes $d(i{,}r)$. Resulting in a $n_T \times h$ dissimilarity matrix that is used to train a classification model. Also for the test set the distance to the prototypes is measured and used for classification. A disadvantage of the k-means is that it cannot handle categorical variables directly. An alternative algorithms by Huang (1997) is available for categorical data. However, an alternative option is to use indicator vector for the categorical features, for example a categorical feature with J categories is vectorised into J features of which each vector contains zero's and one's. A one if the observation belongs to the belonging category and a zero if not.

2.3 Dissimilarity Feature Selection

If the representation set is formulated by using the complete training set of observations the dissimilarity structure is of size $n_T \times n_T$. If prototypes or exemplars are used to formulate the representation set, the dissimilarity structure is of size $n_T \times h$. For both it is sensible to generate an optimal subset of dissimilarity features in such a way that the selected dissimilarity features optimally separate the class labels. As a consequence the amount of noise and the computational time is reduced. In a dissimilarity space, a dissimilarity feature completely isolates the class labels if the pairwise dissimilarities between observations with identical class labels are consistently smaller than the pairwise dissimilarities between observations with different class labels. If we restricted ourselves to logistic regression we might apply a forward stepwise selection procedure based on the Akaike information criterion(AIC) or a L_1 regularised logistic regression (Friedman et al., 2010). But since our aim is to evaluate the performance of FDS classification in combination with several conventional statistical learning techniques we need some universal dissimilarity feature selection method that is applied in combination with a large range of statistical learning techniques.

One suitable universal dissimilarity feature selection method is found by applying the compactness measure (Duin, 1999) as defined in expression 2.1, 2.2 and 2.3. Expression 2.2 allow us to estimate the compactness for each dissimilarity feature and is denoted as a set of $C = \{C_1, C_2, \ldots, C_{n_T}\}$ given a dissimilarity matrix of size $n_T \times n_T$. By applying expression 2.2 a compactness measure is calculated for each dissimilarity feature. This estimate for each dissimilarity feature seems like a good candidate for selecting a subset of dissimilarity features since it estimates how well a specific dissimilarity feature differentiates between the class labels. To select a subset of dissimilarity features by using the compactness measure C we need to define a cutoff value κ that defines a set of dissimilarity features with $C_i \geq \kappa$. For example, if we define the cutoff value κ as $\kappa = 1$ we only select those dissimilarity features which perfectly separate the class labels. Two sensible ways for selecting a cutoff value κ is by taking the average compactness as defined in 2.4 or apply a cross-validation procedure to find the optimal cutoff value for κ. We prefer the latter since it allows us to select an optimal subset of dissimilarity features. The optimal cutoff value for κ is found by applying the following procedure: First we define a set of cutoff values for κ that ranges between 0.5 and the maximum of C. For each cutoff value in this range a 10-fold cross validation procedure is applied to get an estimate of the misclassification rate while using a specific classifier. The cutoff value with the lowest misclassification rate is used to define a subset of dissimilarity features.

To our knowledge, this dissimilarity feature selection method based on the compactness of a dissimilarity structure has never been evaluated and has not been presented in any foregoing scientific literature. In section 2.5 and 2.6 two small experiments are presented with the aim of evaluating the performance of compactness based selection. The first simulation study evaluates the performance of compactness based selection with respect to forward stepwise selection and L_1 regularisation in logistic regression. In section 2.6 compactness based selection is applied in combination with a linear support vector machine and real world data. The performance of compactness based selection is compared to FDS classification without selecting a subset of dissimilarity features and the use of the original feature vector data. However, ahead of these two sections a description is given of how FDS classification is applied in a classifier.

2.4 Feature-Based Dissimilarity Space Classification

For feature-based dissimilarity space(FDS) classification a representation set $R = \{r_1, r_2, \ldots, r_h\}$ that is constructed by using exemplars or prototypes is required to train a specific classifier $f(d)$. The representation set R is constructed by selecting a subset of dissimilarity features in the case of exemplars, in the case of prototypes k-means is used to construct R. In a classification task we require a training set T of size n_T and validation set V of size n_V. To train a classifier in the dissimilarity space a distance function d is required that determines the pairwise dissimilarities between the observations in the training set T and the objects in the representation set R. As a consequence a $n_T \times h$ dissimilarity matrix is formulated by using a distance function $d(i,r)$. Given a classifier $f(d)$ a classification model is trained by applying $f(d(i, r))$ where $i = \{1, \ldots, n_T\}$ and $r = \{1, \ldots, h\}$. $f(d)$ might include traditional classifiers such as linear regression, logistic regression but also Naive Bayes or a support vector machine.

The validation set V is used to evaluate the performance of a trained classifier. Again the pairwise dissimilarities are formulated by the distance function d. The dissimilarity matrix for the validation set is of size $n_V \times h$ and accommodates pairwise dissimilarities between observations in the validation set V and the objects in the representation set R.

2.5 Monte-Carlo Simulation Experiment

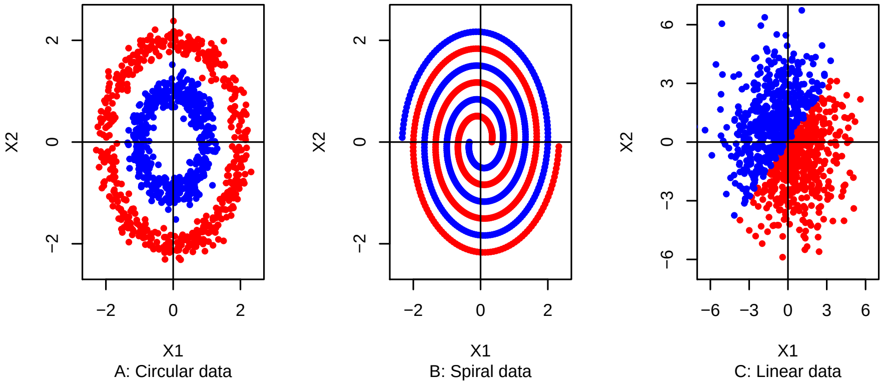
As noted earlier, we conducted a small simulation experiment to evaluate the performance of compactness based selection as defined in the previous paragraph by using expression 2.1, 2.2 and 2.3. In the context of FDS classification compactness based selection is used to select a subset of dissimilarity features. To evaluate the performance of compactness based selection in the context of FDS classification we compared its performance in terms of the misclassification rate with forward stepwise selection and L_1 regularisation (Friedman et al., 2010) that are both applied in the context of FDS classification. We are also interested in comparing the computational time since the computational time of forward stepwise selection and L_1 regularisation drastically gets bigger if n is increasing.

Due to this choice we restricted ourselves to FDS classification applied in a logistic regression model. The three methods are applied over three different datasets with each four different conditions. The three datasets are displayed in Figure 2.2, Figure 2.2A reveals a circular dataset, Figure 2.2B a spiral dataset and Figure 2.2C a linear dataset. Each dataset has four different conditions, a normal condition as displayed in Figure 2.2, a noise condition in which 10% of the class labels are switched, an irrelevant condition in which an irrelevant feature is added and a combination of noise and an irrelevant feature. The irrelevant feature is constructed by using a random uniform distribution $\backsim U(0, 10)$.

The three datasets are selected by their degree of complexity, the linear dataset is characterised as a low complex classification task and is follow by the circular dataset in terms of complexity. The most complex classification task is the spiral data. Each dataset is of size $n = 1000$ and without the irrelevant feature it accommodates two features, therefore each dataset is of size $n \times 2$ and is constructed as a feature vector space. To transform these datasets into a dissimilarity structure we applied the Euclidean distance as defined in expression 2.6.

The current simulation study is a so called Monte-Carlo simulation experiment and is characterised by the concept of generating a new sample from a known distribution. Within each repetition a training and validation set is randomly selected from the generated data. For each condition we used a training set of size 100 or 300, so we have 24 different conditions for each dataset and we have three conditions(selection method) within each simulation. The size of the validation set is dependent on the size of the training set, if the training set is of size 100 the validation set is of size 900. If the training set is of size 300 the validation set is of size 700. The simulation study has 50 replications per condition and in each replication the misclassification rate for each dissimilarity feature selection method is computed and the computational time of each method is measured.

In each replication we randomly select a training set T with n_T observations and a validation set V with n_V observations. Given the Euclidean distance function the dissimilarity matrix for the training set is formulated as $d(i, i)$ and is of size $n_T \times n_T$, it accommodates n_T dissimilarity features. Each dissimilarity feature is described by a vector of distances computed between the observation associated with the dissimilarity feature and the remaining



A: Circular data

B: Spiral data

C: Linear data

Figure 2.2: Overview of used datasets. Each dataset has two class labels, the class labels are represented by the red and blue dots in each dataset. Each dataset has 1000 observations and accommodates two features, X_1 and X_2

n_T observations. All three selection methods are applied over the dissimilarity matrix $d(i, i)$ to construct a representation set $R = \{r_1, r_2, \dots, r_h\}$ of h exemplars and is formulated by one of the selection methods. The final dissimilarity matrix $d(i, r)$ is of size $n_T \times h$ and is used to train the model.

Since all the datasets have a two class outcome variable the focus will be on a two class classification problem. Lets denote π as the probability that the outcome variable Y is in one of the two classes, $\pi = Pr(Y = 1)$. This probability for a specific observation depends on the h dissimilarities features towards the h exemplars, and is denoted as $\pi(d_i) = Pr(Y = 1|d_i)$. As noted earlier, in the current simulation study we focus on the use of logistic regression, thereby we define the probability $\pi(d_i)$ as:

$$\pi(d_i) = \frac{\exp(\alpha + d_i^T \beta)}{1 + \exp(\alpha + d_i^T \beta)}, \tag{2.9}$$

where d_i represents a row vector of dissimilarities between observation i and the h exemplars. β is a vector of h regression coefficients. The optimal value for β is found by minimising the binomial deviance:

$$D = -2 \sum_i y_i \log \pi(d_i) + (1 - y_i) \log(1 - \pi(d_i)). \tag{2.10}$$

For the validation set V a dissimilarity matrix $d(l, r)$ where $l = \{1, \dots, n_v\}$ and $r = \{1, \dots, h\}$. The distance matrix represents the dissimilarities between the observations in the validation set V and the R exemplars as constructed by one of the selection methods.

This dissimilarity matrix is of size $n_V \times h$ and accommodates so called unseen observations. These unseen observations are used to assign a class label to each observation in the validation set V by using the optimised model defined in 2.15. The probability $Pr(Y = 1|d_i)$ for observation i is calculated by using expression 2.14. Where d_i now represents a row vector of observation i from the dissimilarity matrix.

The optimal cutoff value κ that is required for compactness based selection is found by a 10-fold cross validation procedure. For each dataset a vector of possible cutoff points is defined, each vector starts at 0.5 and increases in steps of 0.05 until the maximum, $\max(C)$, of C is reached for that specific dataset. So for each value in the vector of cutoff values a 10-fold cross validation procedure is applied. In the case of L_1 regularisation we also used a 10-fold cross validation procedure to find the optimal penalty parameter. These 10-fold cross validation procedures are applied within in each replication. The optimal cutoff value and penalty parameter with the lowest misclassification rate are selected within each replication. The misclassification rate is defined by dividing the number correct classifications by the total number of classifications.

2.5.1 Results

The results for the linear dataset are displayed in Table 2.1. Overall the results for the linear dataset indicate excellent performance of FDS classification in terms of the misclassification rate. Evidently is the computational time for compactness based selection, Table 2.1 suggests that the in majority of the conditions the computational time for compactness based selection is larger than the computational time of the forward stepwise selection and L_1 regularisation. At first hand the results in Table 2.1 do not suggest any large differences between the selection methods in terms of the misclassification rate. However, the results do indicate that overall L_1 regularisation outperforms compactness based selection and performs similar to forward stepwise selection. The standard deviation of compactness based selection for all four conditions suggest a similar trend compared to the other two selection methods, this implies an equally stable selection method compared to L_1 regularisation and forward stepwise selection.

The results of the circular dataset are displayed in Table 2.2. Overall the results for the circular data again suggest excellent performance of FDS classification in terms of the misclassification rate. Overall the required computational time is larger for compactness based selection than the two other selection methods. However, the difference is less substantial as compared to the linear data. In the noise and irrelevant feature combined condition,

	NORMAL	NOISE	IRRELEVANT	n&IR
C - n=100	0.030 (0.015,3)	0.065 (0.026,5.9)	0.049 (0.02,2)	0.095 (0.033,1.5)
L_1 - n=100	0.026 (0.015,0.3)	**0.052** (0.021,0.4)	0.032 (0.014,0.2)	**0.075** (0.028,0.4)
Stepwise - n=100	**0.019** (0.009,1)	0.073 (0.039,1.6)	**0.028** (0.013,1)	0.101 (0.044,2.1)
C - n=300	0.018 (0.008,24.1)	0.041 (0.013,15.7)	0.022 (0.009,15.3)	0.058 (0.014,9.5)
L_1 - n=300	0.013 (0.008,0.5)	**0.032** (0.014,1.9)	0.013 (0.006,0.3)	**0.042** (0.013,0.9)
Stepwise - n=300	**0.008** (0.005,5.2)	0.072 (0.04,53.2)	**0.010** (0.006,7.3)	0.077 (0.048,43.5)

Table 2.1: Results linear data: Selection method and the training set size are displayed in the left column. The values in the table reveal the average misclassification rate, within brackets the standard deviation of the misclassification rate and the computational time(seconds) are displayed.

compactness based selection is much faster than L_1 regularisation and forward stepwise selection. Noteworthy is the misclassification rate of forward stepwise selection in the noise and irrelevant combined condition, the results in Table 2.2 suggest that the misclassification rate and the computational time is substantially increased. For a training set of size 300 the computational time of forward stepwise selection is on average 109.6 seconds for the noise and irrelevant combined condition. The results also indicate that L_1 regularisation is less sensitive to noise and irrelevant features. Again the results indicate that compactness based selection is equally stable as the other two methods since the standard deviations are almost equal. For the noise and irrelevant combined condition it seems that forward stepwise selections becomes more unstable since the standard deviation increases. Noteworthy is that the circular data is generally characterised as a more complex classification task as compared to the linear data. However, the observed misclassification rate is lower in the circular data as the observed misclassification rate in the linear data.

For the spiral data the results are displayed in Table 2.3. Given a training set of size 100 the performance of FDS classification is far from excellent. However, in the normal condition compactness based selection and L_1 regularisation evidently outperform forward stepwise selection. In each condition with a training set of size 100, forward stepwise selection seems hardly superior to randomly selecting a class label. As soon as noise or an irrelevant feature is added all the methods barely perform better than chance. Given a normal condition and a training set of size 300 the results in Table 2.3 suggest that compactness based selection outperforms(0.15) the L_1 regularisation(0.32) and forward stepwise selection(0.388). Although the misclassification rate has increased in the noise condition, the performance of compactness based selection is still superior(0.302) compared to the L_1(0.42) and forward stepwise selection(0.441). In the condition where an irrelevant feature is added, or in combination with noise, the performance of all three methods is barely better than chance. Interesting to see is that the required computational time for compactness based selection is less with respect to L_1 and forward stepwise selection in all conditions.

	NORMAL	NOISE	IRRELEVANT	n&IR
C - n=100	**0.003** (0.002,1.8)	0.035 (0.034,1.2)	**0.039** (0.029,0.7)	0.154 (0.056,0.6)
L_1 - n=100	0.007 (0.01,0.2)	**0.011** (0.011,0.3)	0.053 (0.023,0.7)	**0.091** (0.035,1)
Stepwise - n=100	**0.003** (0.003,0.9)	0.025 (0.035,0.9)	0.169 (0.168,1.4)	0.289 (0.15,1.9)
C - n=300	**0.001** (0.002,13.1)	0.01 (0.01,7.5)	**0.009** (0.006,4.6)	0.086 (0.039,3.4)
L_1 - n=300	0.003 (0.002,0.4)	**0.004** (0.003,0.6)	0.02 (0.007,3.4)	**0.028** (0.014,4.4)
Stepwise - n=300	0.002 (0.002,5.4)	0.012 (0.025,11.2)	0.022 (0.009,13)	0.152 (0.1,109.6)

Table 2.2: Results circular data: Selection method and training set size are displayed in the left column. The values in the table reveal the average misclassification rate, within brackets the standard deviation of the misclassification rate and the computational time(seconds) are displayed.

	NORMAL	NOISE	IRRELEVANT	n&IR
C - n=100	0.409 (0.047,0.5)	**0.446** (0.029,0.5)	**0.492** (0.019,0.4)	**0.493** (0.018,0.5)
L_1 - n=100	**0.397** (0.052,2.2)	0.46 (0.034,1.8)	0.494 (0.015,2)	0.5 (0.017,2.2)
Stepwise - n=100	0.493 (0.032,0.7)	0.491 (0.024,0.6)	0.497 (0.014,0.6)	0.497 (0.016,0.4)
C - n=300	**0.15** (0.04,2.5)	**0.302** (0.043,2.6)	**0.473** (0.019,2.6)	**0.485** (0.019,2.7)
L_1 - n=300	0.32 (0.06,13.5)	0.42 (0.037,8.7)	0.492 (0.021,6.5)	0.493 (0.019,8.5)
Stepwise - n=300	0.388 (0.151,42)	0.441 (0.083,52.8)	0.495 (0.018,3.1)	0.497 (0.017,6.8)

Table 2.3: Results spiral data: Selection method and the training set size are displayed in the left column. The values in the table reveal the average misclassification rate, within brackets the standard deviation of the misclassification rate and the computational time (seconds) are displayed.

2.5.2 Conclusion

A thorough inspection of the results suggest that in some specific conditions compactness based selection evidently outperforms forward stepwise selection and L_1 regularisation in terms of the misclassification rate. The most substantial difference in performance between compactness based selection and the other two selection methods is found in Table 2.3, given a training set of size 300, compactness based selection evidently outperforms the L_1 regularisation and forward stepwise selection for the normal and noise condition in the spiral data. Similar results were not found in the linear and circular data. Nonetheless, the results in Table 2.2 of the circular data suggest that the performance of compactness based selection is comparable to L_1 regularisation and outperforms forward stepwise selection.

After transforming the data from a feature vector space into a dissimilarity space we computed the compactness/complexity measure(as defined in expression 2.1) for each dataset in the normal condition. The linear dataset after transformation revealed a compactness measure of 0.75, the circular dataset revealed a compactness measure of 0.58 and the spiral dataset revealed a compactness measure of 0.50. This compactness measure for each dataset is the average compactness measure over 50 Monte-Carlo simulations. This compactness measure is an indicator of the complexity of a classification task. The compactness measures indicate that the spiral data is the clearly the most complex classification task and is followed by the circular data in terms of complexity. This in combination with the results in Table 2.1, 2.2 and 2.3 indicates that in the case of a complex classification task compactness based selection outperforms forward stepwise selection and L_1 regularisation.

Additionally, the results also suggest that the performance of FDS classification is more beneficial in combination with a complex classification task. The performance of FDS classification in Table 2.1 with respect to the performance of FDS classification in Table 2.2 suggest that the performance in the circular data outperforms the linear data in each condition and selection method. Noteworthy is that the linear data is a less complex classification task compared to the circular data. This indicates that FDS classification is more beneficial in the case of a more complex classification task.

The difference in the complexity also explains the decrease in computational time for the compactness based selection method, as noted earlier we applied a 10-fold cross validation procedure to find an optimal cutoff value. For each dataset we generated a vector of cutoff values that started with 0.5 and increased in steps of 0.05 until $\max(C)$ and performed a 10-fold cross validation procedure for each value in vector of cutoff values. However, $\max(C)$ is much lower for a highly complex dataset with respect to a less complex dataset. As a consequence, the size of the vector of cutoff values between 0.5 and $\max(C)$ is much smaller for a highly complex dataset. Therefore, the number of 10-fold cross validations is reduced and as a consequence the average computational time is lower. Noteworthy is the fact that we used standardised R functions for L_1 regularisation (Friedman et al., 2010) and forward stepwise selection (R Core Team, 2014). A function in R was written for the compactness based selection method. In contrast to compactness based selection the L_1 regularisation and forward stepwise selection are highly optimised functions. Due to the optimisation of these functions it is very likely that these functions require less computational time.

Overall we conclude that the use of compactness based selection for selecting a subset of dissimilarity features is superior to L_1 regularisation and forward stepwise selection in the case of a complex classification task. In a low complex classification task, such as the linear dataset L_1 regularisation is preferred over compactness based selection. However, the results also suggest that FDS classification is not beneficial in the case of a low complex classification task.

2.6 Support Vector Machine - Bootstrap Study

Our second assessment of compactness based selection is characterised by a so-called bootstrap procedure and involves the use of linear support vector machines(SVM) (Vapnik, 1996). These SVM's are applied on two distinctive datasets in combination with FDS classification. The aim of a linear SVM is to separate the class labels by an optimal separating hyperplane. This hyperplane is defined by a decision boundary which maximizes its margin around the decision boundary in such a way that it separates the class labels optimally. As a result it maximizes the distance to the closest points from both classes by using so-called support points. These are the points that are defined to be on the boundary of the margin. However, this concept does not hold if the cases are non-separable by a linear decision boundary. Therefore Vapnik (1996) generalised the concept of an optimal separating hyperplane to the non-separable case. In the case that the classes overlap in the feature space the aim of linear SVM is still to maximise the margin around the decision boundary, but now allow some observations to be inside the margin around the decision boundary. To find an optimal separating hyperplane while some observations are allowed to be inside the margin around the decision boundary a cost parameter ω is required. The cost parameter ω controls the number of observations within the margin of the decision boundary. Normally the optimal value for ω is found by a cross validation procedure, in this small bootstrap study we used an arbitrary chosen value for ω that is equal to 1. For a more thorough description about the computations and algorithms of linear SVM we refer to Hastie et al. (2009).

Similar to the previous simulation study our aim is to asses the performance of compactness based selection in combination with FDS classification. The performance of compactness based selection is assessed by inspecting the average misclassification rate obtained while classifying unseen observations. To obtain an estimate of the average misclassification rate a bootstrap procedure is used. By means of this procedure the additional value of compactness based selection in the context of linear SVM's is evaluated.

A bootstrap procedure is characterised by the concept of resampling with replacement from a given dataset to generate a bootstrap sample. This bootstrap sample is used as a training set, the observations outside the bootstrap sample are used as a validation set. In our situation we are interested in the misclassification rate. All the models are fitted in R by using the package e1071 (Meyer et al., 2015). The data consists of two distinctive datasets:

1. The first dataset is the well known Iris flower data (Fisher, 1936) and has four continuous features that represents the length and the width of the sepals and petals for each observation. The data accommodates $n = 150$ observations and has 50 observations for each of the three flower species. In the classification task our aim is to predict the correct flower species. The outcome variable accommodates three class labels, Iris setosa, Iris virginica and Iris versicolor. The complexity of the Iris data is estimated at 0.93, this indicates a low complex classification task. The complexity is estimated by standardising all the original features and by using a dissimilarity matrix of size $n \times n$ and is constructed by the Euclidean distance function. Each feature is standardized by subtracting the features mean value and dividing by the features standard deviation.

2. The second dataset is the Bupa data and has six continuous features and $n = 345$ observations. The first five features are blood tests which are thought to be sensitive to liver disorders. The last feature measures the alcohol consumption of each individual observation in the data. The aim of the data is to predict if an individual has a liver disorder. The outcome variable has two labels, liver disorder(145) and no liver disorder(200). The complexity of the Bupa data is estimated at 0.507. This indicates a highly complex classification task. The complexity is estimated in a similar manner as estimated for the Iris data.

For each dataset we have four different conditions: a condition in which compactness based selection is used to select a subset of dissimilarity features, these dissimilarity features are used to train a linear SVM. In the second condition is FDS classification is applied in the context of a linear SVM without selecting a subset of dissimilarity features. In the third condition the original feature vector data is used to train a SVM. The last condition uses a randomly selected subset of dissimilarities features to train a model. The number of randomly selected dissimilarity features is optimised by a 10-fold cross validation procedure.

The bootstrap procedure is identical for each condition and dataset, however, within each bootstrap replication each condition requires additional optimisations. We limited the number of bootstrap replicates to $B = 100$. Each bootstrap procedure is defined by the following steps:

1. For each bootstrap replicate b we generate a bootstrap sample T of size n_T from the data. The bootstrap sample is generated by sampling with replacement from the data and represents the training set for the b^{th} bootstrap replicate. For the Iris data $n_T = 100$ and for the Bupa data $n_T = 200$.

2. Within each bootstrap replication a validation set V of size n_V is generated. The validation set V accommodates all the observations that are not observed in the bootstrap sample T. For the Iris data the validation set V is on average of size $n_V = (1 - 1/n)^{n_T} \cdot n = 77$. For the Bupa data the validation set V is on average of size $n_V = 137$ (Efron and Tibshirani, 1993).

3. In the case of FDS classification each feature is standardized by subtracting the features mean value and dividing by the features mean absolute deviation. The training set T is transformed in a dissimilarity matrix D_{tr} by using the Euclidean distance and is used to train a classifier. The validation set V is also transformed in a dissimilarity matrix D_{vl} by using the Euclidean distance and represents the validation set.

4. Train a classifier by using T or D_{tr}.

5. Classify the unseen observations in V or D_{vl} by using the trained classifier.

6. Calculate an estimate of the misclassification rate er_b for the b^{th} bootstrap replicate.

Some conditions require an additionally procedure between step three and four: In the case of compactness based selection a 10-fold cross validation is used between step three and four to find the optimal cutoff parameter κ. For each dataset a vector of possible cutoff points is defined, each vector starts at 0.5 and increases in steps of 0.01 for the Bupa data and 0.05 for the Iris data until $\max(C)$ is reached for that specific dataset.

In the case of randomly selecting a subset of dissimilarity features the optimal number of randomly selected dissimilarity features is found by a 10-fold cross validation procedure between step three and four. The optimal number of randomly selected dissimilarity features is found between 1 and 30 for the Iris data and between 1 and 80 for the Bupa data.

The misclassification rate is estimated by using the validation set V of size n_V. The whole procedure is 100 times repeated and the average of the 100 misclassification rates for each condition was taken as an estimate of the overall misclassification rate for each specific condition.

For the Bupa data we also evaluated the misclassification rate as a function of the training set size n_T. The results are obtained by repeating the bootstrap procedure as described above for several different training set sizes n_T. We started with a training set of size $n_T = 20$ and increased in steps of 5 units until a training set size of $n_T = 300$ is reached. The size of validation set V was kept constant at $n_V = 45$ by randomly selecting a subset of 45 observations from the validation set V that accommodates the observations that are not observed in the bootstrap sample T.

The bootstrap procedure for estimating the misclassification rate as introduced above differs from the original bootstrap procedure as introduced by Efron and Tibshirani (1993). Given a dataset of size n a bootstrap sample is created by sampling with replacement n observations from the original data. In the original bootstrap procedure for misclassification rate estimation this bootstrap sample of size n is used to train a classifier and the observations that are not sampled into the bootstrap sample are used as a validation set. A consequence of the bootstrap procedure is that the number of dissimilarity features in FDS classification is equal to n. Due to the large number of dissimilarity features the computational time is drastically increased for training a classifier. To reduce the computational time a bootstrap sample of size n_T is taken. A consequence of adjusting the bootstrap sample size is that the estimate of the misclassification rate is high biased. This is due to the fact that some observations in the bootstrap sample are sampled multiple times and as a consequence the number of unique observation is reduced in the bootstrap sample (Efron, 2004). In the original bootstrap procedure as introduced by Efron and Tibshirani (1993) a bias correction is presented, however, this bias correction does not apply to a bootstrap procedure with an adjusted bootstrap sample of size n_T.

2.6.1 Results

The results are displayed in Table 2.4 and for the Iris data the results suggest that SVM's with the original feature vector data outperforms FDS classification with and without compactness based selection. For the Iris data the misclassification rate of FDS classification(0.06) is equal to FDS classification in combination with compactness based selection(0.06). The misclassification rate for the original feature vector data is 0.04 and is 0.02 lower with respect to the other three conditions. The misclassification rates for the Iris data are within one standard deviation, so in terms of the misclassification rate we cannot determine a superior technique. However, the number of wins per technique are displayed in Table 2.4 and represents the number of times a specific classifier had the lowest classification rate. The total number of wins do not sum up to 100 since two techniques might have the lowest misclassification rate simultaneously. Table 2.4 reveals that SVM in combination with the original feature vector data managed to win 91 times. However, a large number of these wins are shared with FDS classification since the total number of wins for the Iris data sum up to 176 instead of 100. Noteworthy is that we did not observe a difference between the FDS classification techniques. The results indicate that the misclassification rate of FDS classification is independent of the selection of dissimilarity features.

	Iris		Bupa	
	Error	Wins	Error	Wins
FDSC	0.06(0.04)	27	0.31(0.04)	26
FDSC & C	0.06(0.03)	29	**0.29**(0.03)	57
FDSC & R	0.06(0.04)	29	0.31(0.03)	32
OFV	**0.04**(0.03)	91	0.34(0.03)	7

Table 2.4: Results bootstrap for each condition and dataset: FDSC & C is the misclassification rate for FDS classification in combination with compactness based selection. FDSC is the misclassification rate for FDS classification, OFV is the misclassification rate while using the original feature vector data. FDS & R is the random selection of dissimilarity features. Within the brackets the standard deviation of the misclassification rate is presented.

The results for the Bupa data in Table 2.4 suggest that FDS classification in combination with compactness based selection(0.29) outperforms the linear SVM with the original feature vector data(0.34). The results in Table 2.4 suggest that FDS classification in combination with compactness based selection slightly outperforms the other FDS classification techniques. The results in table 2.4 with regard to FDS classification are within one standard deviation. Randomly selecting a subset of dissimilarity features in FDS classification performs(0.31) similar to normal FDS classification. The difference between FDS classification in combination with compactness based selection and the traditional SVM in terms of the average misclassification rate is about 0.05. The average decrease in the misclassification rate while using FDS classification in combination with compactness based selection is 0.02 with respect to normal FDS classification. The number of wins also suggest a similar trend, FDS classification with compactness based selection managed to get the lowest misclassification rate 57 times out of 100 replications. The number of wins for normal FDS classification and the random selection of dissimilarity features is almost equal. Similar to the Iris data the number of wins do not sum up to 100, this is due to fact that two methods might obtain an identical misclassification rate.

For the Bupa data we also evaluated the misclassification rate as a function of the training set size. The results are obtained by repeating the bootstrap study as described in section 2.6 for several different training set sizes. We started with a training set of size 20 and increased the size in steps of 5 units until a training set of size 300 is reached. The results are displayed in Figure 2.3. The results suggest that for a small training set ($n < 75$) the performance of FDSC and the OFV condition is better than the performance of FDSC & C and FDSC & R . However, if the training set becomes larger ($n > 75$) FDSC & C is outperforming the OFV condition. When the training set is larger than 150 the difference between FDSC & C and OFV becomes constant and is around 0.035. The results also indicate that FDSC & C outperforms FDSC and FDSC & R when $n > 125$.

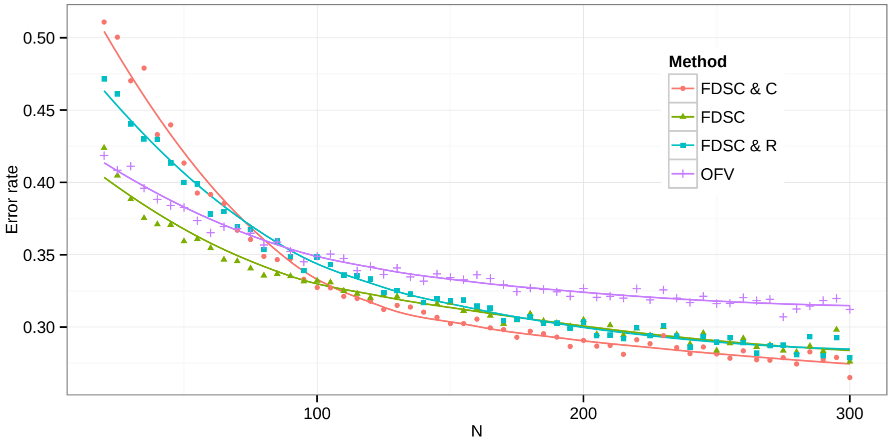


Figure 2.3: The misclassification rate as a function of training set size. FDSC & C is FDS classification in combination with compactness based selection. FDSC & R is the random selection of dissimilarity features. FDS is feature dissimilarity space classification and OFV is SVM while using the original feature vector data.

2.6.2 Conclusion

The results of the current bootstrap procedure suggest a similar trend as observed in the Monte-Carlo simulation study described in the previous section. The results suggest that FDS classification is superior in the context of a complex classification task. For the Bupa data the results, as displayed in Table 2.4, suggest that the use of compactness based selection is beneficial in terms of the number of wins as compared to FDS classification applied without any dissimilarity feature selection method. In terms of the misclassification rate the superiority of FDS classification in combination with compactness based selection with respect to the other FDS classification techniques cannot be determined. For the Iris data, the performance of FDS classification is almost identical to FDS classification in combination with compactness based selection. Despite the fact that these two methods almost perform identical the latter method is preferred since it reduces the number of dissimilarity features. A reduced number of dissimilarity features is mainly beneficial in terms of the computational effort. Nonetheless, the use of compactness based selection is only preferred if the training set is large. For a small training set Figure 2.3 indicates that the FDS classification is preferred in terms of the misclassification rate. In the case of a small training set the compactness measure C for each dissimilarity feature is based on a small sample of observations. As a consequence the estimate of C for each dissimilarity feature is unstable and hardly to be generalised to the validation set. In the case of the Bupa data FDS classification in combination with compactness based selection was beneficial if the training set is of size 100 or larger. The required training set size is dependent on the complexity of a classification task. For a less complex classification task it is likely that the training set requires less samples until compactness based selection becomes beneficial.

The results of the Iris data suggest that the use of a linear SVM in combination with the original feature vector data is superior to all FDS classification methods. In combination with the contrasting results of the Bupa and Iris data the results in Table 2.4 suggest that FDS classification is not beneficial in the context of a low complex classification task. A low complex classification task is often characterised by data that is optimally separated by a linear decision boundary. The Iris dataset is well known and is proven to be easily separated by a linear decision boundary. In the preceding Monte-Carlo simulation study the linear data as displayed in Figure 2.2C is also optimally separated by a linear decision boundary. Nonetheless, FDS classification does not optimally separate the data by a linear decision boundary but uses a flexible decision boundary. As a consequence FDS classification is more sensitive to overfitting in the case where a linear decision boundary is preferred. An important note is that the compactness measure used for estimating the complexity of a classification task is dependent on the distance function. In the current study we used the Euclidean distance, however, changing the distance function will result in a different estimate for the complexity of a classification task.

3 Data

In the final part of this thesis we will assess the performance of feature-based dissimilarity space (FDS) classification while applying it to a practical problem. The data primarily accommodates bio-markers which are believed to be related with the prevalence of depressive disorders. Patient data were obtained from the Netherlands Study of Depression and Anxiety (NESDA, Penninx et al. 2008). NESDA is an ongoing longitudinal cohort study on the longterm development and consequences of depression and anxiety. NESDA data are collected by multiple centres and baseline measurement are obtained for 2981 patients and control subjects. For a detailed description we refer to Penninx et al. (2008).

Our aim is to predict the presence of a recurrent or persistent depressive disorder in unseen observations. A recurrent depressive disorder is characterised by intervals of periods with and without depression. Individuals with a persistent depressive disorder are characterised by the presence of a depression over at least a period of two years. A depression is defined by a depressed mood that is present for a large part of the day and almost every day. Individuals experience a loss of interest and pleasure in activities that are normally experienced as pleasurable. Additional symptoms of a depression might differ between individuals and ranges between cognitive restrictions, such as a diminished ability to think, concentrate and indecisiveness, and physical restrictions such as a decreased level of energy, appetite changes and a change in psychomotor activities.

The outcome variable has two class labels (disorder and no disorder). The disorder class label refers to individuals with a recurrent or persistent depressive disorder. No selection has been applied on the basis of comorbidity, so the disorder class label includes individuals with and without comorbidity. Of the 2981 observations, 982 observations had missing values. All observations with missing values are removed from the analysis. The no disorder label refers to individuals without a recurrent or persistent depressive disorder but may include different disorders such as a single depressive episode or anxiety disorder. There are 705 individuals with a recurrent or persistent depression and 1294 individuals without a persistent or recurring depressive disorder. The presence of a persistent or recurrent depression was assessed by using a standardised psychiatric interview (CIDI)(Kessler et al., 2004).

Additionally to the outcome label an additional outcome label is specified. This additional outcome label is based on the idea that individuals with a single depressive episode at baseline may develop a persistent or recurrent depression over time. So, for all the individuals with a single depressive episode at baseline we evaluated if some of these individuals were diagnosed with a persistent or recurrent depression three years after the first diagnoses was established. An example of this process is depicted in Table 3.1 where PRD_B refers to a persistent or recurrent depression at baseline and PRD_3 refers to a persistent or recurrent depression three years after baseline. SD_B refers to a single depressive disorder at baseline. Y_{or} refers to the original outcome label and Y_{add} to the additional outcome label. Y_{add} is identical to Y_{or} for the majority of observations. Y_{add} and Y_{or} only differ for those observations which are diagnosed with a persistent or recurrent depression three years after baseline and where diagnosed with a single depressive episode at baseline. An example of this is depicted in the first row of Table 3.1.

The theory behind this is that individuals with a single depressive episode may have the same biological parameters as individuals with a persistent or recurrent depression which may indicate that these individuals eventually develop a persistent or recurrent depression. We identified 54 (2.7%) individuals which are diagnosed with a single depressive episode at baseline and are diagnosed with a recurrent or persistent depression three years after baseline. In the set of original outcome labels these 54 individuals are labeled as no disorder, for the additional outcome labels these 54 individuals are changed into the disorder label. n remains of size $n = 1999$.

PRD_B	PRD_3	SD_B	$\dots$	Y_{or}	Y_{ad}
0	1	1	$\dots$	0	1
1	1	0	$\dots$	1	1
0	0	1	$\dots$	0	0
0	1	0	$\dots$	0	0

Table 3.1: Example formulation additional outcome label. PRD refers to a persistent or recurrent depression and SD refers to a single depressive episode. B refers to baseline and 3 refers to three years after baseline.

3.1 Features

The 12 features used for training and prediction are selected on the basis of scientific literature that describe the possible relation between depressive disorders and the specific feature.

Gender The data accommodates 687 males and 1312 females. Out of the 687 males 238 were diagnosed with a recurrent or persistent depression. 467 females were diagnoses with a persistent or recurrent depression.

Age The average age is 42.1(Sd = 12.9) years. The average age for individuals with a persistent or recurrent depression is 43.6(Sd = 11.8), for individuals without a persistent or recurrent depression the average age is 41.3(Sd = 13.3).

Vitamin D25 In the recent years vitamin D25 has been increasingly linked to cognitive deterioration (Annwiler et al., 2009) and psychiatric disorders (Cherniack et al., 2009). A recent study revealed that a low level of vitamin D25 is associated with the presence of depressive disorder (Milaneschi et al., 2014). Vitamin D metabolites influence the growth of neurones by the up-regulation of nerve growth factors (Neveu et al., 1994). The average level of D25 over all individuals is 63.72 (Sd = 29.02). For individuals with a recurrent or persistent depression the average level of vitamin D25 is 61.42 (Sd = 26.41). For individuals without a recurrent or persistent depression the average level of vitamin D25 is 64.99 (Sd = 30.29).

Brain-Derived Neurotrophic Factor Brain-derived neurotrophic factor(BDNF) is a nerve growth factor that is found in the brain and is responsible for the survival and growth of neurones in the central nervous system (Acheson et al., 1994). BDNF is active in brain regions that are related to learning, memory and higher thinking, these brain regions include the hippocampus, cortex, and the basel forebrain (Yamada and Nabeshima, 2003). Research has revealed that exposure to stress reduces the expression of BDNF. If the expression of BDNF is persistently reduced in the related brain regions it will eventually result in atrophy of these related brain regions. Atrophy of these related brain regions is observed in individuals with a depressive disorder (Warner-Schmidt and Duman, 2006). In our data the average expression of BDNF is 9.25 (Sd = 3.46), for individuals with a recurrent or persistent depression the average expression of BDNF is 8.78 (Sd = 3.04). For individuals without a recurrent or persistent depression the average expression of BDNF is 9.50 (Sd = 3.65).

High Sensitive C-reactive Protein High sensitive C-reactive protein (hs-CRP) is an inflammation marker and it is argued that stress, deprivation and other negative cognitions trigger and inflammatory response that is similar to a general inflammatory response in the case of a bodily disease. An inflammatory response sets the brain into a status of sickness and often the symptoms of being physically sick overlap with the symptoms of having a depressive disorder (Berk et al., 2013). The average expression of hs-CRP is 2.90 (Sd = 5.11). For individuals with a recurrent or persistent depression the average expression of hs-CRP is 2.61 (Sd = 4.75). For individuals without a recurrent or persistent depression the average expression of hs-CRP is 3.06 (Sd = 5.29).

Interleukine-6 Similar as the previous feature Interleukine-6(IL-6) is also an inflammatory marker and is related to an inflammatory response of the body in case physical distress. It is argued that IL-6 mediates the expression of BDNF in related brain regions (Sharma et al., 2008). This assumes that an inflammatory response induces the expression of BDNF. The average expression of IL-6 is 1.35 (Sd = 3.89). For individuals with a recurrent or persistent depression the average expression of IL-6 is 1.07 (Sd = 2.38). For individuals without a recurrent or persistent depression the average expression of IL-6 is 1.50 (Sd = 4.50).

Tumour Necrosis Factor Alpha Tumour necrosis factor alpha (TNFa) is also an inflammatory marker and is related to depression in a similar way as hs-CRP. Dowlati et al. (2010) revealed that the expression of TNFa is elevated in depressed individuals. Our data suggest a contrary and very small difference between depressed individuals and controls. This difference is likely due to the inclusion criteria. The average expression of TNFa is 1.13 (Sd = 1.51). For individuals with a recurrent or persistent depression the average expression of TNFa is 1.07 (Sd = 1.35). For individuals without a recurrent or persistent depression the average expression of TNFa is 1.17 (Sd = 1.59).

Creatinine Creatinine is an indicator of renal functioning and several studies revealed an elevated expression of creatinine in individuals with a depressive disorder (Allen, 2012; Segal et al., 2007). The exact relation between depression and creatinine is still unknown. However, it is argued that creatinine is involved in the energy metabolism in neurotransmission which is alternated in depressive individuals. The average expression of creatinine is 81.11 (Sd = 14.37). For individuals with a recurrent or persistent depression the average expression of creatinine is 90.40 (Sd = 11.83). For individuals without a recurrent or persistent depression the average expression of creatinine is 76.04 (Sd = 13.04).

Aspartate Transaminase Aspartate transaminase is an indicator of liver functioning. The exact relation with depression is unknown, however, it is reasonable to argue that an induced expression of aspartate transaminase in depressed individuals is due to the use of anti-depressants. The average expression of aspartate transaminase is 26.06 (Sd = 10.99). For individuals with a recurrent or persistent depression the average expression of aspartate transaminase is 22.75 (Sd = 10.87). For individuals without a recurrent or persistent depression the average expression of aspartate transaminase is 27.87 (Sd = 10.63).

Glucose Level The glucose level is related to diabetes. Individuals with diabetes are more vulnerable to a depressive disorder (Anderson et al., 2001). The average level of glucose is 5.17 (Sd = 1.01). For individuals with a recurrent or persistent depression the average level of glucose is 5.30 (Sd = 0.85). For individuals without a recurrent or persistent depression the average level of glucose is 5.10 (Sd = 1.09).

Cholesterol The relation between cholesterol levels and depression has been extensively studied and several inconsistent results has been published (Tanskanen et al., 2000). The average level of cholesterol is 1.63 (Sd = 0.45). For individuals with a recurrent or persistent depression the average level of cholesterol is 1.65 (Sd = 0.44). For individuals without a recurrent or persistent depression the average level of cholesterol is 1.61 (Sd = 0.45).

WHO Disability Assessment Schedule WHO Disability Assessment Schedule is a questionnaire that measure the degree of disability and is the only non-biological measure. It is generally assumed that the degree of disability is higher for individuals with a recurrent or persistent depression in comparison to individuals without a recurrent or persistent depression. However, our data suggest no difference in the degree of disability between the two groups. The average score is 24.95 (Sd = 21.16). For individuals with a recurrent or persistent depression the average score is 24.28 (Sd = 21.07). For individuals without a recurrent or persistent depression the average score is 25.32 (Sd = 21.20).

3.2 Formulation Dissimilarity Matrix

In the case of FDS classification the original feature vector data is transformed into a dissimilarity space. The Euclidean distance is used to construct a dissimilarity structure. For the feature that represents gender an indicator vector is used (0 = male, 1 = female). All the features are standardised before applying the distance function. Each feature is standardized by subtracting the features mean value and dividing by the features standard deviation. The final dissimilarity matrix D is of size $n \times n$. The matrix D is split into a training matrix D_{tr} of size $n_T \times n_T$, and a validation matrix D_{vl} of size $n_V \times n_T$.

4 Methods

In this section we will give a detailed description of how the performance of feature-based dissimilarity space(FDS) classification is evaluated. Our aim is to assess the performance of FDS classification in terms of the discrimination power and the misclassification rate for several techniques which includes Logistic regression, Support Vector Machine (SVM), Naive Bayes, Random forest and linear/quadratic discriminant analysis.

The discrimination power of a classifier is evaluated by estimating the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. The AUC is represented by a single number and a graphical representation in which the true positive rate is plotted against the false positive rate (1-specificity). The advantage of the AUC with respect to the misclassification rate is that the AUC is insensitive to an imbalanced set of class labels. Basically the AUC quantifies the overall ability of a classifier to discriminate between two different class labels at different threshold values. A classifier that classifies at random will obtain an AUC of 0.5, a perfect classifier will obtain an AUC of 1 (Fawcett, 2006).

To obtain an accurate estimate of the misclassification rate and the AUC for each classification technique and condition we applied a bootstrap procedure with B, $B = 100$, bootstrap replications. Due to the computational effort that is required in some classification techniques we limited the number of bootstrap replications to $B = 100$. The design of the bootstrap procedure is identical to the bootstrap procedure discussed in section 2.6. However, in the current bootstrap procedure the training set T (bootstrap sample) in step one is of size $n_T = 500$. The size of n_T is limited to 500 due to the computational efforts in FDS classification. Due to this the validation set is on average of size $n_V = (1-1/n)^{n_T} \cdot n = 1557$. In step six the AUC and the misclassification rate is estimated in each bootstrap replicate.

Additional steps are required between the third and the fourth step (as described in section 2.6) for some specific classification techniques and conditions. Since FDS classification is capable of being applied in several conventional classification techniques we will assess the performance of FDS classification by using a diverse set of different classification techniques and conditions. We defined a total of five different conditions which are characterised by:

OFV Classification by using the original feature vector (OFV) data: A classifier is trained in each bootstrap replicate by using the training set T of size n_T. The misclassification rate and AUC are estimated by using the validation set V.

FDSC FDS classification: A classifier is trained by using the dissimilarity matrix D_{tr} and the misclassification rate and AUC are estimated by using the dissimilarity matrix

D_{vl}. In this condition the dissimilarity matrix D_{tr} is of size $n_T \times n_T$, the validation matrix D_{vl} is of size $n_V \times n_T$. The validation and training matrix are formulated in each bootstrap replicate.

FDSC & C FDS classification in combination with compactness based selection: In this condition a classifier is trained by using a subset of dissimilarity features which are selected by using compactness based selection. The optimal cutoff value for compactness based selection is estimated by using a 10-fold cross validation procedure within each bootstrap procedure. The 10-fold cross validation procedure is applied between third and the fourth step. The 10-fold cross validation procedure is applied on the training matrix D_{tr} as defined in each bootstrap replication. A classifier is trained in each bootstrap replicate by using the selected subset of dissimilarity features.

FDSC & L_1 FDS classification in combination with L_1 regularisation: In this condition a classifier is trained by using a subset of dissimilarity features which are selected by using L_1 regularisation (Friedman et al., 2010). The penalty parameter for L_1 regularisation is estimated by using a 10-fold cross validation procedure within each bootstrap procedure. The 10-fold cross validation procedure is applied between third and the fourth step. The 10-fold cross validation procedure is applied on the training matrix D_{tr} as defined in each bootstrap replication. A classifier is trained by using the selected subset of dissimilarity features.

FDSC & P FDS classification in combination with prototypes: The prototypes are formulated by using k-means (Macqueen, 1967) as described in section 2.1.1. For each class label a set of 60 prototypes is created in each bootstrap replicate. A set of 250 random initial starting value was used to prevent a local minima. The prototypes are created by using the training set T. The prototypes are created between the third and the fourth step. The training matrix D_{tr} is of size $n_T \times h$. The validation matrix D_{vl} is of size $n_V \times h$.

The aim is to compare the OFV condition with the remaining conditions. So, the aim is to evaluate the performance of FDS classification with respect to the condition in which the original feature vector data is used to classify unseen observations. However, due to the principle of bootstrapping each condition is not applied in each technique. An overview of which condition is applied in which technique is presented in Table 4.1.

Noteworthy in Table 4.1 is radial basis SVM (Vapnik, 1996), Radial Basis SVM is only applied in combination with the original feature vector data. This is due to the nature

	Condition				
Method	OFV	FDSC	FDSC & C	FDSC & L_1	FDSC & P
Logistic Regression	√	√	√	√	
Linear SVM	√	√	√		
Radial Basis SVM	√				
Naive Bayes	√	√	√		√
Random Forests	√	√	√		√
Linear Discriminant Analysis	√	√	√		√
Quadratic Discriminant Analysis	√				√

Table 4.1: Condition per method: Quick overview of each condition and in which classification technique it is applied.

of the radial basis kernel function which also uses the Euclidean distance. By taking a combination of radial basis SVM and FDS classification the radial basis kernel function will formulate similarities over already formulated dissimilarities. Since both techniques use a distance function it implies that these two techniques strongly overlap. FDS classification could also be interpreted as a kernel function. A kernel function is denoted by $K(x_{ip}, x_{rp})$, the radial basis kernel is defined by:

$$K(x_{ip}, x_{rp}) = \exp(-\gamma ||x_{ip} - x_{rp}||^2), \tag{4.1}$$

where x_{ip} and x_{rp} represents object i and r with a representation on feature p. $||x_{ip} - x_{rp}||^2$ can be recognised as the Euclidean distance. The γ parameter is optimised by a 10-fold cross-validation procedure within each bootstrap replication. For a more detailed description about how kernels are applied in the context of a SVM classifier we refer to Hastie et al. (2009). In the case of linear SVM the cost parameter ω is also optimised by a 10-fold cross validation procedure within each bootstrap procedure.

Also noteworthy in Table 4.1 is the QDA which is only applied in combination with the original feature vector data and FDS classification in combination with prototypes. Applying FDS classification in the context of QDA without selecting a subset of dissimilarity features is unachievable since the the covariance matrix for each class label is rank deficient. A matrix is said to be rank deficient if $p \geq n$. In the case of FDS classification $n = p$. As a consequence the covariance matrix is not positive definite which is a necessity to obtain an inverse of the covariance matrix. A possible solution is to use compactness based selection to reduce the number of dissimilarity features to satisfy the assumption of $p < n$. However, due to the nature of a bootstrap simulation it is possible that the training set accommodates identical observations. As a consequence the dissimilarity matrix D_{tr} accommodates a number of non-unique dissimilarity features. In this case the covariance matrix is also rank deficient. To overcome the problem of obtaining identical dissimilarity features prototypes are formulated as discussed above for each class label.

Random Forests is a classification technique that operates by formulating a user defined number of de-correlated classification trees during training and classifies by using the majority vote of all these classification trees (Breiman, 2001). For the construction of a random forest model we used the R package randomForest (Liaw and Wiener, 2002). Each individual tree is constructed by the principle of bagging and random selection of features. For each individual classification tree a bootstrap sample is taken from the training set T and a constant number of random features is selected (Ho, 1998). The number of random selected features in each classification tree is $\sqrt{p}$ for each classification technique and condition.

Naive Bayes is a probabilistic classifier that uses Bayes theorem to classify unseen observation. The term naive originates from the assumption that all features are independent in a Naive Bayes classifier. A Naive Bayes classifier aims to assign unseen observations to a class label by using the probability of a class label given the data. Given a set of class labels $G_l = \{G_1, \ldots, G_L\}$ and a data matrix X with p features and n observations in the original feature vector space Naive Bayes aims to assign a probability to each G_l given X, $P(G_l | x_{n1}, \ldots, x_{np})$. This classification method works fine while using categorical features, however, in practice continuous features are common. For all continuous features in the original feature vector space and the dissimilarity space we assume a normal distribution and the Gaussian probability density function is used to estimate the probability for each class label.

because our aim is to assess the performance of FDS classification within each individual technique by comparing it with the same classification technique while using the original feature vector data as input. The performance of each condition and technique is assessed by evaluating the average misclassification rate and the average AUC. The average misclassification rate is estimated over $B = 100$ bootstrap replications and is defined by the following expression:

$$\widetilde{er} = \frac{1}{B} \sum_{b=1}^{B} er_b. \tag{4.2}$$

The average AUC is formulated in a similar manner and is defined by the following expression:

$$\widetilde{auc} = \frac{1}{B} \sum_{b=1}^{B} auc_b. \tag{4.3}$$

By taking the average misclassification rate and the average AUC these estimates are more robust against strong deviations in these estimates than single estimates. A density distribution curve for the misclassification rate is also presented for each technique and condition. This density distribution curve provides an indication of the spread of the misclassification rate. For the AUC a ROC curve is presented for each technique and condition, these curves indicate to what extend a classifier is capable in differentiating between the class labels over different threshold values. The ROC curve is estimated over the 100 bootstrap replications by taking the class probabilities for the unseen observations in each bootstrap replicate.

Additionally to the bootstrap procedure a model independent importance measure is presented to evaluate the importance of each individual feature in the original feature vector space. This method uses a permutation approach to evaluate the importance of each original feature. Before transforming the original feature vector data into a dissimilarity matrix a feature of interest is permutated. After permutation the dissimilarity matrix is formulated and the compactness is estimated by using expression 2.4. A low compactness measure indicates a more complex classification problem. The importance measure is formulated as the compactness measure of the original data minus the compactness measure of the permuted data. This permutation approach is repeated 100 times for each feature and is applied on the complete dataset. This technique allows us to estimate the importance of an individual feature in a model independent manner.

For the model with the best performance in terms of the misclassification rate and the AUC an additional misclassification rate will be estimated by using the additional outcome label as discussed in section 3. Additionally for the model with the best performance the sensitivity and specificity is discussed. The sensitivity and specificity for the model with the best performance is based on the 100 bootstrap replications. The results of the best model are also discussed in terms of the clinical implications.

5 Results

In this section the results of the bootstrap study for each technique and condition are presented. The results are divided in two subsections. In the first subsection, the results with regard to the comparison between feature-based dissimilarity space (FDS) classification and traditional classification are presented. In the second subsection the results with regard to the clinical implications are given. In the second subsection section the model with the best performance in terms of the misclassification rate and AUC is thoroughly discussed. Additionally the importance of each original feature as observed in the original feature vector space is discussed.

5.1 Results Feature-Based Dissimilarity Space Classification

In this section the results of each technique and condition is presented per classification technique. Within each classification technique the comparison between the traditional technique and the technique in combination with FDS classification is discussed. The traditional technique is characterised by the use of the original feature vector data. The classification techniques are compared by evaluating the misclassification rate and the AUC.

5.1.1 Logistic Regression

The results of the bootstrap study with regard to logistic regression and FDS classification are presented in Table 5.1. The results suggest a poor performance of FDS classification while applied without selecting a subset of dissimilarity features. The average misclassification rate of FDS classification is 0.31, this is barely better than chance since 705 (35%) individuals are diagnosed with a recurrent or persistent depression in the original sample. The standard deviation of the misclassification rate also suggest a more unstable classifier as compared to the other methods. In Figure 5.1 the distribution of the misclassification rate for each technique is displayed. The distribution of the misclassification rate of FDS classification (without selecting subset of dissimilarity features) clearly reveals that the misclassification rate is much more unstable and widely spread as compared to the other techniques. These results are a consequence of the fact that logistic regression is incapable of finding an optimal estimate for each coefficient when the amount of features is equal to the amount of observations. It is reasonable to assume that FDS classification without selecting a subset of dissimilarity features performs similar to the random selection of outcome labels.

Table 5.1 reveals that traditional logistic regression applied while using the original feature vector data performs worse when comparing it with FDS classification in combination with selecting a subset of dissimilarity features. The average misclassification rate of logistic regression is 0.19. Over 100 bootstrap replications logistic regression only managed to get the lowest misclassification rate five times. FDS classification in combination with compactness based selection has an average misclassification rate of 0.18 and managed to get the lowest misclassification rate 12 times. The total number of wins does sum up to 100 in this case. However, since it is possible that two techniques simultaneously obtain the lowest misclassification rate the total could also deviate from 100.

	error	AUC	wins
Log Reg FDS & C	0.18(0.014)	0.88	12.00
Log Reg FDS & L_1	0.17(0.013)	0.90	83.00
Log Reg FDS	0.31(0.097)	0.71	0.00
Log Reg	0.19(0.014)	0.86	5.00

Table 5.1: Results bootstrap study logistic regression: Traditional logistic regression (log) compared to FDS classification. C refers to compactness based selection and L_1 to L_1 regularisation. Wins refers to the number of times the classification technique obtained the lowest misclassification rate. The standard deviation of the misclassification rate is displayed within the brackets.

FDS classification in combination with L_1 regularisation outperformed all other methods in terms of the misclassification rate and the AUC. The average misclassification rate of FDS classification in combination with L_1 regularisation is 0.17 and managed to win 83 times. On average FDS classification in combination with L_1 regularisation is 2% better than traditional logistic regression. The standard deviation of 0.013 reveals a stable method, this is confirmed by inspecting Figure 5.1.

In Table 5.1 the area under the cure (AUC) of the ROC curve is presented for each technique. The AUC of FDS classification without selecting a subset of dissimilarity features has the worst AUC, 0.71, and is poorly capable of discriminating between the class labels. FDS classification in combination with L_1 regularisation is most capable of discriminating between the class labels and has an average AUC of 0.90. In Figure 5.2 it is clearly visible that the ROC curve associated with FDS classification in combination with L_1 regularisation has the largest distance to the black diagonal line. The average AUC of FDS classification in combination with compactness based selection is 0.88 and the ROC curve as displayed in figure 5.2 is close to the curve of FDS classification in combination with L_1 regularisation and traditional logistic regression. FDS classification in combination with compactness based selection is slightly better than traditional logistic regression in discriminating between the class labels.

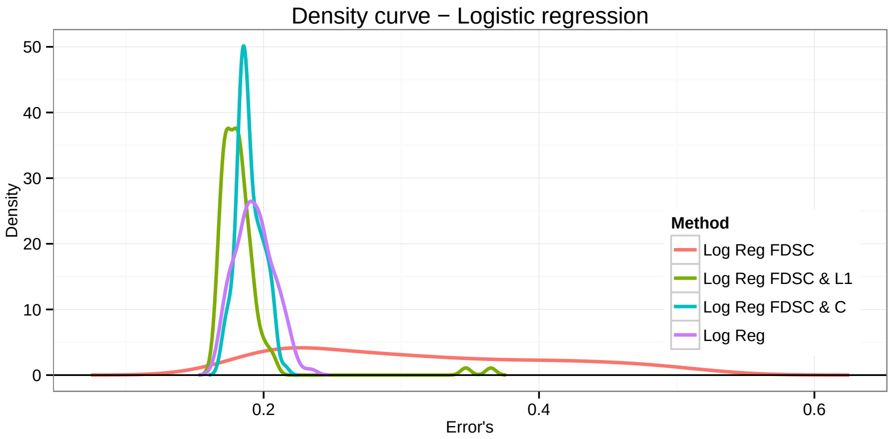


Figure 5.1: The distribution of the misclassification rate for each technique estimated over 100 bootstrap replications.

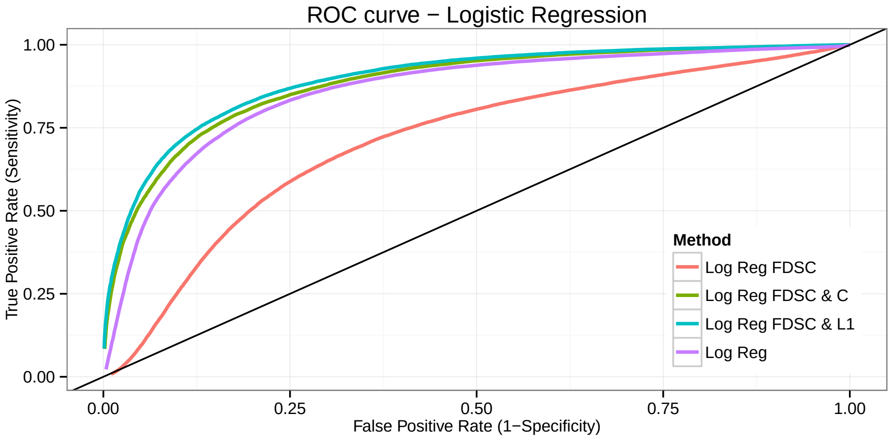


Figure 5.2: Receiver operating characteristic curve (ROC) for each technique. A ROC curve that is close to the black diagonal line is often labeled as a classifier with poor performance.

5.1.2 Support Vector Machine

The results of the bootstrap study with regard to support vector machine's (SVM) are presented in Table 5.2. The results indicate that FDS classification is a strong competitor of radial basis SVM. The average misclassification rate of FDS classification with and without selecting a subset of dissimilarity features is 0.17, together these two methods managed to get the lowest misclassification rate 56 times out of the 100 bootstrap replications. Radial basis SVM managed to get the lowest classification rate 54 times over 100 bootstrap replications, the average misclassification rate for radial basis SVM is 0.17. The only difference between FDS classification and radial basis SVM is observed in the AUC, The AUC of radial basis SVM is 0.90 and 0.89 for both FDS classification techniques.

	error	AUC	wins
L-SVM FDS	0.17(0.012)	0.89	22.00
L-SVM FDS & C	0.17(0.008)	0.89	34.00
L-SVM	0.19(0.014)	0.87	2.00
RB-SVM	0.17(0.012)	0.90	54.00

Table 5.2: Results bootstrap study. Traditional linear SVM (L-SVM) compared to FDS classification and radial basis (RB) SVM. C refers to compactness based selection. Wins refer to the number of times a classification technique had the lowest misclassification rate. The standard deviation of the misclassification rate is displayed within the brackets.

Inspection of the density curve in Figure 5.3 reveals that the density curves of FDS classification largely overlaps with the density curve of radial basis SVM. A close inspection of these curves reveal that the density curve of the radial basis SVM is slightly more to the left. The density curve of the traditional linear SVM and FDS classification has substantially less overlap. This indicates the superiority of FDS classification in terms of the misclassification rate. Noteworthy is that the average performance of the linear SVM in terms of the misclassification rate and the AUC is equal to the performance of logistic regression while using the original feature vector data, Table 5.1. A similar trend is seen when FDS classification is applied in the context of a linear SVM, the average misclassification rate of FDS classification in combination with compactness based selection is 0.17. FDS classification in combination with compactness based selection and logistic regression has an average misclassification rate of 0.18 and 0.17 in combination with L_1 regularisation.

The AUC as presented in Table 5.2 indicates a similar trend as compared to the misclassification rate. FDS classification and radial basis SVM are superior to a linear SVM while using the original feature vector data. Radial basis SVM is slightly more capable of discriminating between the class labels as compared to the FDS classification methods. Figure 5.4 indicates that ROC curves for FDS and radial basis are almost identical, linear support vector machine clearly performs less as compared to FDS classification and radial basis.

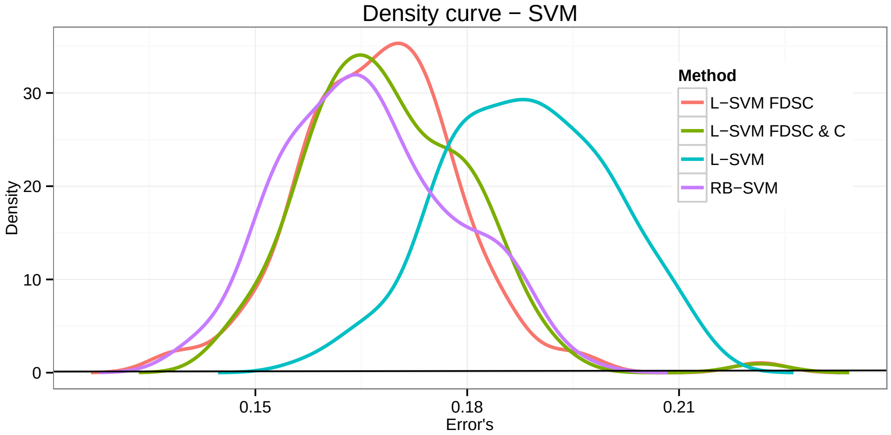


Figure 5.3: The distribution of the misclassification rate for each technique estimated over 100 bootstrap replications.

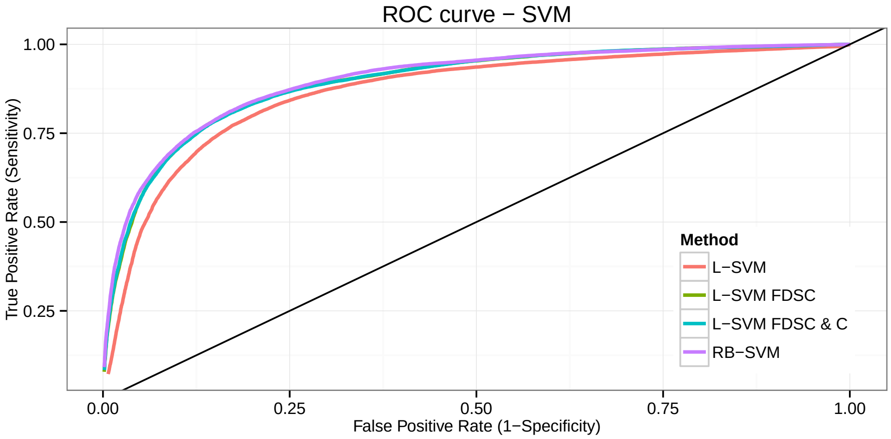


Figure 5.4: Receiver operating characteristic curve (ROC) for each technique. An ROC curve that is close to the black diagonal line often represents poor performance.

5.1.3 Naive Bayes Classifier

The results in Table 5.3 suggests a poor performance of Naive Bayes in each condition. The performance of Naive Bayes while using the original feature vector data is superior to all FDS classification methods. The average misclassification rate is 0.34 while using the original feature vector data. This is barely better than randomly selecting a class label. Figure 5.6

indicates that the distribution of the misclassification rate is widely spread for each technique and indicates that each technique has a poor performance in terms of the misclassification rate and is worse than randomly selecting a class label. The poor performance of Naive Bayes is reflected in the AUC, the AUC of all FDS classification techniques indicate that Naive Bayes in combination with FDS classification is incapable of selecting the correct class label for an unseen observation. The ROC curve in Figure 5.5 reveals a similar trend and indicates that the original Naive Bayes classifier is the best choice.

	error	AUC	wins
N-B FDS	0.57(0.100)	0.67	0.00
N-B FDS & C	0.50(0.125)	0.61	3.00
N-B FDS & P	0.38(0.056)	0.68	18.00
N-B	0.34(0.069)	0.79	79.00

Table 5.3: Results bootstrap study. Naive Bayes (N-B) compared with FDS classification. C refers to compactness based selection and P to the use of prototypes. The amount of wins refer to the amount of times a classification technique had the lowest misclassification rate. The standard deviation of the misclassification rate is displayed within the brackets.

Inspection of the original Naive Bayes classifier reveals that each unseen observation is assigned to the non-depressive class label. The likelihood of the data given a class label, $p(X|G_l)$, as estimated by a gaussian distribution for each continuous feature is identical for each class label. As a consequence the classification of unseen observations solely depends on the prior probability $p(G_l)$. The dependence on the prior probability indicates that Naive Bayes in combination with a gaussian density function is incapable of finding a pattern in the data that differentiates between the class labels.

Inspection of the Naive Bayes classifiers in combination with FDS classification suggest that due to the large amount of probability estimates for the dissimilarity features FDS classification is incapable of finding a stable pattern in the data. This is likely due to the combination of a large number of dissimilarity features and the insensitivity of Naive Bayes to detect a pattern in the data. The insensitivity to detect a pattern is likely due to the assumption that continuous features are normally distributed. Inspection of the dissimilarity features revealed that these are non-normally distributed.

Inspection of Figure 5.6 also reveals that FDS classification with and without selecting a subset of dissimilarity features either assigns all unseen observations to the depressive or the non-depressive class label since the two peaks are centred around 0.35 and 0.65.

FDS classification in combination with prototypes seems more stable than the other FDS classification techniques. It is likely that this is due to the fact that this technique only uses 120 dissimilarity features.

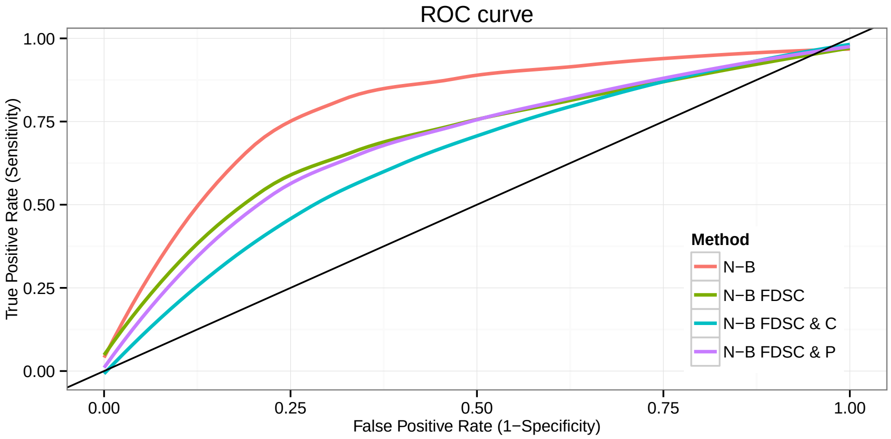


Figure 5.5: Receiver operating characteristic curve (ROC) for each technique. A ROC curve that is close to the black diagonal line represents poor performance of a classifier.

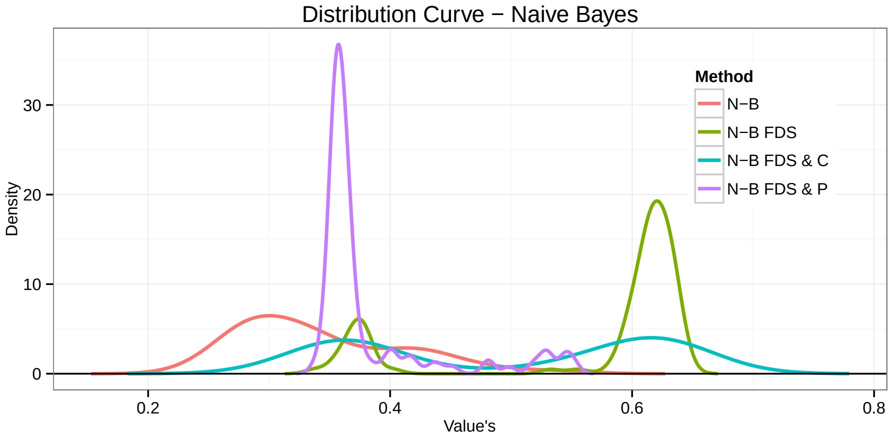


Figure 5.6: The distribution of the misclassification rate for each technique estimated over 100 bootstrap replications.

5.1.4 Random Forest

The results of the bootstrap simulation in which FDS classification is applied in the context of Random Forests are displayed in Table 5.4. The results clearly reveal that FDS classification is not beneficial in the context of a Random Forests classifier. The average misclassification rate of FDS classification with and without selecting a subset of dissimilarity features is 0.16. Similar results are found while using the distance between observations and prototypes. The average misclassification rate of the traditional Random Forests model is 0.14 and is 0.02 lower than FDS classification. Out of the 100 bootstrap replications the traditional random forest model managed to get the lowest misclassification rate 93 times. The difference in the misclassification rate is clearly visible in Figure 5.7, the overlap in terms of the misclassification rate between the traditional Random Forests model and FDS classification is minimal and are almost separated. These results indicate that the traditional Random Forests model is superior to FDS classification in terms of the misclassification rate. 14 out of 100 unseen observations are misclassified while using the traditional random forest model.

	error	AUC	wins
R-F FDS	0.16(0.013)	0.90	6.00
R-F FDS & C	0.16(0.012)	0.90	2.00
R-F FDS & P	0.16(0.012)	0.90	0.00
R-F	0.14(0.011)	0.93	93.00

Table 5.4: Results bootstrap study. Random forests compared with FDS classification. C refers to compactness based selection, P to the use of prototypes. The amount of wins refer to the amount of times a classification technique had the lowest misclassification rate. The standard deviation of the misclassification rate is displayed within the brackets.

A similar trend of superiority is revealed in Figure 5.8 that represents the ROC curve for each individual condition. The curves clearly indicate that the discrimination power of the traditional Random Forests model is superior with regard to the FDS classification techniques. The average AUC of the traditional Random Forests model is 0.93. The AUC of each individual FDS classification technique is on average 0.90. These results indicate that the traditional Random Forests model, which uses the original feature vector data, is more valuable in discriminating between class labels.

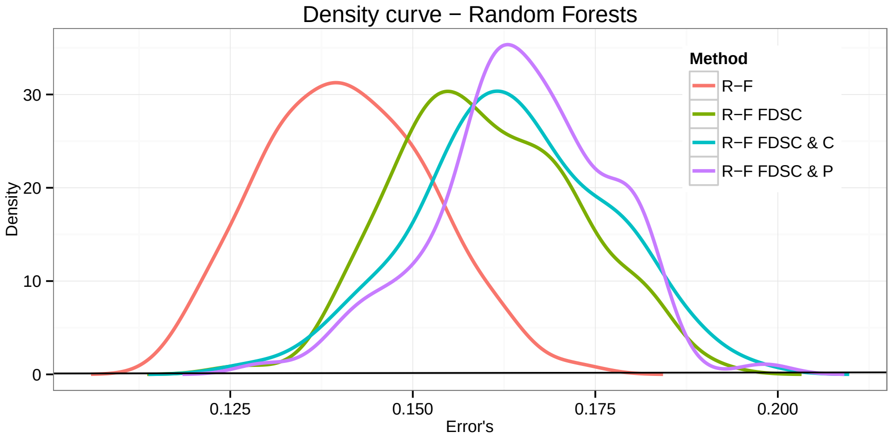


Figure 5.7: The distribution of the misclassification rate for each technique estimated over 100 bootstrap replications.

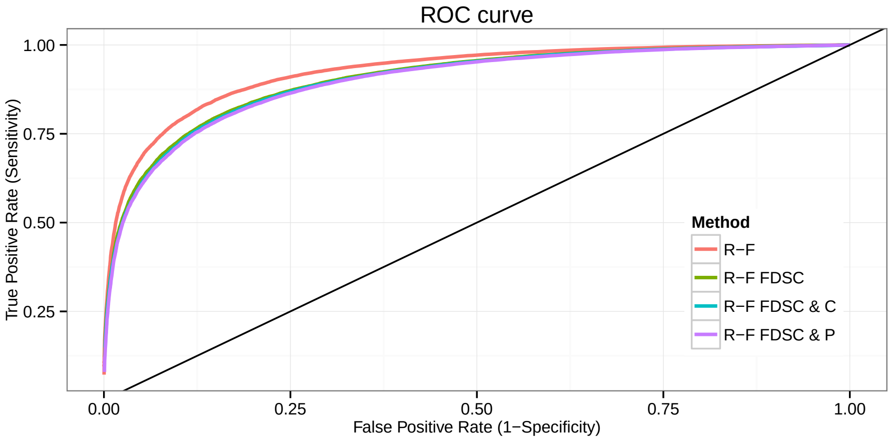


Figure 5.8: Receiver operating characteristic curve (ROC) for each technique. A ROC curve that is close to the black diagonal line represents poor performance of a classifier.

5.1.5 Linear / Quadratic Discriminant Analysis

The results of the bootstrap simulation in which FDS classification is applied in the context of linear and quadratic discriminant analysis are displayed in Table 5.5. The misclassification rates in Table 5.5 suggest that FDS classification, applied in a LDA classifier, is preferred while classifying unseen observations. FDS classification in combination with a LDA classifier managed to obtain the lowest misclassification rate 99 times. Figure 5.9 reveals that the density curves of the misclassification rate are almost identical for the two

FDS techniques which uses exemplars. The average misclassification rate of these techniques is 0.19 and is on average 0.01 better than traditional LDA which uses the original feature vector data.

The worst performance in terms of the misclassification rate is observed in the traditional QDA which uses the original feature vector data. The average misclassification rate of QDA is 0.28, the standard deviation of the misclassification rate and Figure 5.9 indicate an unstable classifier. It is likely that the QDA classifier is unstable due to the violation of the normality and continues features assumption. The original feature vector data contains binary data for gender, this feature is not continuous and not normally distributed. QDA and LDA are not suited for the use of a binary features, as a consequence the covariance matrix for each class label might be unstable.

	error	AUC	wins
LDA FDS	0.19(0.008)	0.87	48.00
LDA FDS & C	0.19(0.010)	0.87	48.00
LDA FDS & P	0.20(0.010)	0.85	3.00
LDA	0.20(0.009)	0.86	2.00
QDA FDS & P	0.24(0.020)	0.80	0.00
QDA	0.28(0.057)	0.81	0.00

Table 5.5: Results bootstrap study. Traditional linear/quadratic discriminant analysis compared with FDS classification. C refers to compactness based selection and P to the use of prototypes. The amount of wins refer to the amount of times a classification technique had the lowest misclassification rate. The standard deviation of the misclassification rate is displayed within the brackets.

Table 5.5 also indicates that the performance of FDS classification in combination with compactness based selection in terms of the misclassification rate outperforms traditional LDA. The average misclassification rate of FDS classification in combination with compactness based selection is 0.19 and is 0.01 better than traditional LDA. The performance of FDS classification in the context of a LDA while using prototypes is similar to traditional LDA in terms of the misclassification rate.

Figure 5.10 and Table 5.5 Indicates that in terms of the AUC FDS classification while using exemplars is superior to all other techniques. Overall the AUC reveals that each techniques is reasonably capable of discriminating between the class labels. The worst performance is observed while using the traditional QDA classifier and QDA in combination with FDS classification and a set of prototypes.

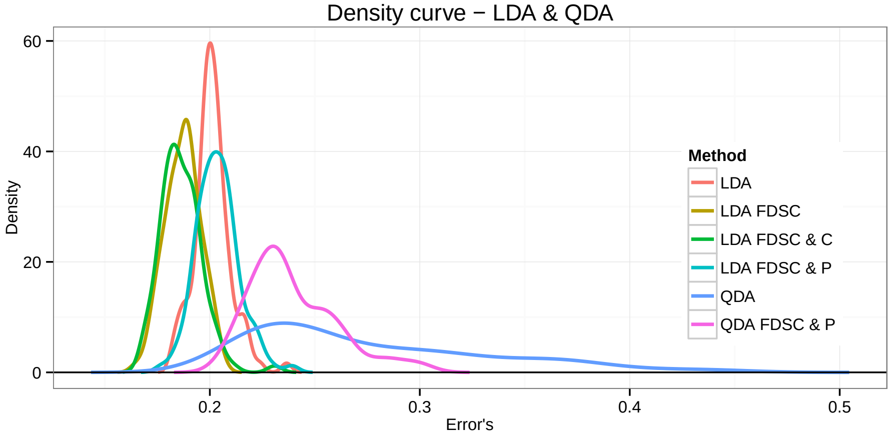


Figure 5.9: The distribution of the misclassification rate for each technique estimated over 100 bootstrap replications.

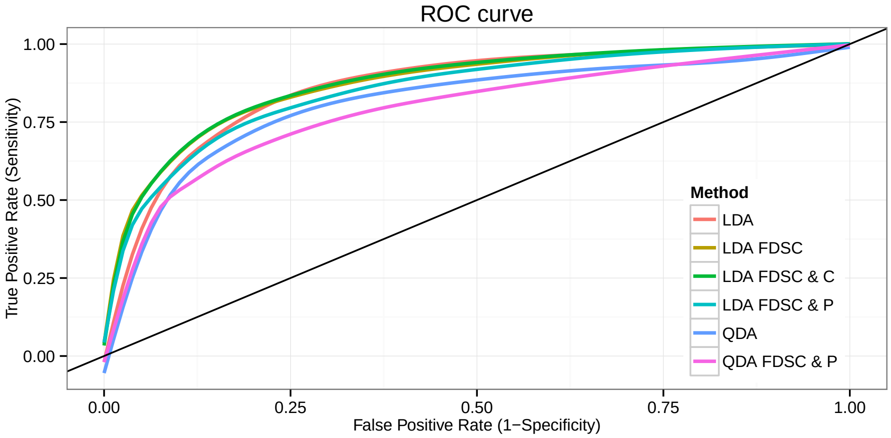


Figure 5.10: Receiver operating characteristic curve (ROC) for each technique. An ROC curve that is close to the black diagonal line often has a poor performance.

5.2 Clinical Results

The best performance is achieved by using a Random Forests classifier in combination with the original feature vector data. The average misclassification rate of this model is 14%. This indicates that out of 100 unseen observations a random forest classifier will on average misclassify 14 observations. For the Random Forests classifier we also estimated the sensitivity and specificity, sensitivity represents the proportion of individuals with a persistent or recurrent depression and are correctly identified as such. Specificity is the proportion of individuals without a persistent or recurrent depression and are identified as such by the random forest model. The sensitivity of the random forest model is 79% and the specificity is 93%. This indicates that the random forest model is better in correctly identifying non-depressive individuals.

The average misclassification rate of the additional outcome label while using the original random forest model is 16.4%. On average this is 2.4% higher than the normal misclassification rate. Since 2.7% of the individuals are diagnosed with a single depressive episode at baseline and are diagnosed with a recurrent or persistent depression three years after baseline the additional misclassification rate suggest a decline in performance in terms of the misclassification rate. This suggests that the pattern of biomarkers that are related to the presence of a recurrent or persistent depressive disorder are a consequence of the presence of a recurrent and persistent depression.

The advantage of using a random forest classifier in combination with the original feature vector data is the capability of assessing the importance of each individual feature in the original feature vector space. The importance of each feature is displayed in Figure 5.11 and reveals that the glucose, aspartate transaminase (ASAT) and creatinine expression are the most important features. The average decrease in accuracy of the random forest classifier after permuting creatinine is estimated at 12.8%. The average decrease in accuracy for aspartate transaminase is estimated at 5.5% and 2.5% for the glucose level. The importance measure for a Random Forests classifier is obtained by applying a permutation approach in each individual classification tree and measures the decrease in accuracy. The average over all trees is used as an importance measure for a single Random Forest model (Liaw and Wiener, 2002). The results are obtained by taking the average decrease in accuracy over all the Random Forests classification models that are in the bootstrap procedure.

We also constructed a model independent importance measure by using a permutation approach in combination with the compactness measure as described in expression 2.4. A decrease in compactness is interpreted as an increase in complexity. These results (Figure 5.12) indicate a similar trend in the top three features as the importance plot obtained by the Random Forests models. The results indicate that creatinine, ASAT and the glucose expression are the most important features that influence the compactness of this classification task. The average decrease in compactness after permutation of the feature creatinine is 0.03. For ASAT the average decrease in compactness is 0.01 and 0.003 for the glucose expression. This importance measure gives an indication to what extend an individual feature contributes to the separation of the class labels in the dissimilarity space. The remaining features are around zero and indicate that they have no impact on the complexity of the classification problem. Although the top three features indicate a similar trend as the importance measure of the Random Forests model, the model independent importance measure also reveals some contrasting trends as compared to the importance measure of the Random Forests model.

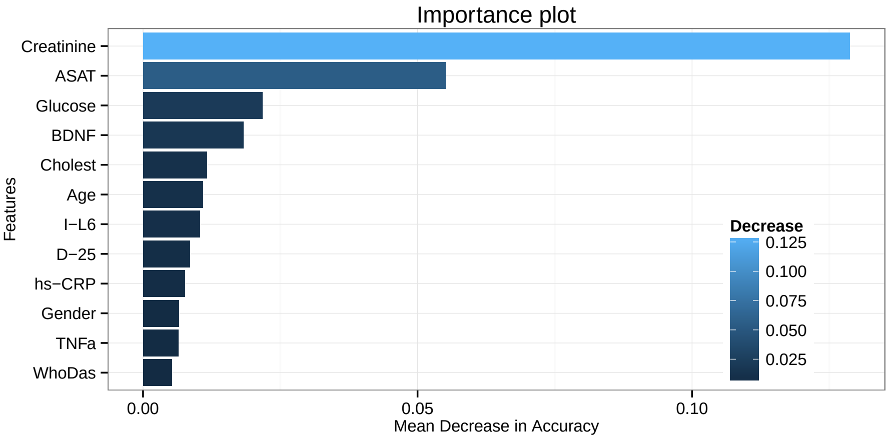


Figure 5.11: Importance of each feature. For each feature the average decrease in accuracy is plotted. The average decrease in accuracy is obtained by taking the average decrease in accuracy over 100 Random Forests classifiers.

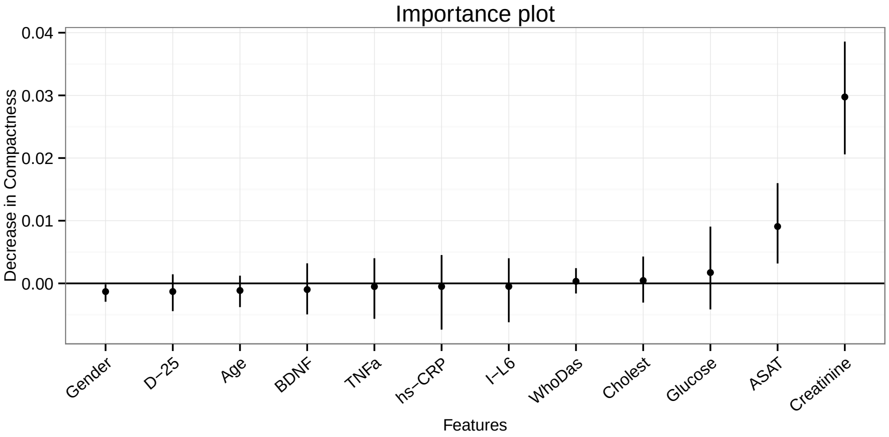


Figure 5.12: Independent measure of importance based on the compactness measure. The middle points represent the average decrease in compactness and the lines represent the 95% confidence interval of the average decrease in compactness for each feature.

6 Conclusion

The aim of this thesis was to asses the performance of feature-based dissimilarity space (FDS) classification by applying it to a practical problem and to interpret the results with respect to the practical problem. Additionally this thesis evaluates how and when FDS classification is beneficially applied. Similar to the previous section this section is divided in two individual subsections. In the first subsection the performance of FDS classification is discussed and an answer on the question when FDS classification is beneficial is given. In the second subsection the results are discussed in terms of the clinical implications.

6.1 Conclusion Feature-Based Dissimilarity Space Classification

The main aim of this thesis was to evaluate the performance of FDS classification with respect to traditional learning algorithms which uses the original feature vector space.

The results in section 5 indicate that in the combination of a complex classification task and a linear classifier the performance of FDS classification is slightly superior to the use of the original feature vector data. The complexity of the data as used in section 5 is estimated at 0.53, this indicates a complex classification task. We used several models that are characterised by a linear decision boundary and are applied in combination with FDS classification in the bootstrap study. The diverse set of linear classifiers includes logistic regression, linear discriminant analysis and linear support vector machine. In all these classification techniques FDS classification slightly outperforms the traditional method in which the original feature vector data is used to train a classifier. These linear classification techniques are fitting a linear decision boundary in the dissimilarity space, this decision boundary is non-linear and flexible in the original feature vector space.

Identical results as observed in section 5 are observed by Duin et al. (2010), in which the performance of FDS classification is evaluated over 301 distinctive datasets and by applying FDS classification to a wide range of classifiers. In their paper FDS classification is evaluated by comparing its performance in terms of misclassification rate with the performance of conventional classifiers. Over 301 datasets the results indicate that FDS classification is mainly beneficial in combination with a linear classifier. Similar results are observed in section 5 where FDS classification was mainly beneficial in combination with linear classifiers. The results of Duin et al. (2010) also indicate that the majority of the 301 datasets are likely to be identified as a complex classification task.

Similar results are found in section 2.5 and 2.6. In section 2.5 compactness based selection is evaluated and a similar trend is observed. The results in section 2.5 indicate that the performance of FDS classification in combination with a linear classifier is more beneficial when the complexity of classification task is increasing. This relation between complexity and performance is observed while inspecting the results of the linear and circular data in Table 2.1 and 2.2. The results reveal that FDS classification in combination with compactness based selection has a lower misclassification rate in the circular data as in the linear data. However, the linear data is identified as a less complex classification task as compared to the circular data. In section 2.6 a similar trend is observed. The results in section 2.6 indicate that FDS classification is mainly beneficial in the context of a complex classification task and a linear classifier.

It is likely that a complex classification task is characterised by a flexible decision boundary while a low complex classification task is characterised by a linear decision boundary. So, in the case of a low complex classification task, such as displayed in Figure 2.2C or the Iris data, a linear classifier such as a fitted by logistic regression is preferred. However, if FDS classification is applied in a classifier such as logistic regression the decision boundary becomes non-linear in the original feature vector space while a linear decision boundary is preferred in a low complex classification task. As a consequence of using a flexible decision boundary while a linear decision boundary is preferred is that the classifier becomes more sensitive to overfitting. Overfitting occurs when a classifier describes noise instead of the underlying relationship. In general an overfitted model will have poor performance in terms of the misclassification rate.

In section 2 compactness based selection is presented as a newly proposed method for selecting a subset of dissimilarity features. This method is derived from the compactness hypothesis (Arkadev and Braverman, 1966) which gives the fundamentals for estimating the complexity of a classification task. In this thesis we generalised this concept of compactness to individual dissimilarity features which basically estimates the compactness for each individual dissimilarity feature. By using the compactness measure for each individual dissimilarity feature it is possible to select a subset of dissimilarity features that separates the class labels in the dissimilarity space. Additionally compactness based selection eliminates noise, dissimilarity features that barely separates the class labels could be considered as random noise. The dissimilarity features that barely separates the class labels are removed form the set of dissimilarity features used to train a classifier.

In section 2.5 and 2.6 the performance of compactness based selection is evaluated in two different settings. The first experiment evaluates the performance of compactness based selection in the context of logistic regression and is compared to L_1 regularisation and forward stepwise selection. All three conditions are applied while using dissimilarity data. The results of this experiment evidently suggest a superior performance of compactness based selection in combination with highly complex data. The complexity is estimated by using the relaxation of the compactness hypothesis as proposed by Duin (1999). In the case of less complex data compactness based selection performs similar to L_1 regularisation and forward stepwise selection is superior in the case of low complex data.

Similar results are found in the experiment as discussed in section 2.6. In this section compactness based selection is applied in the context of linear SVM's while using real world data. In this experiment the performance of compactness based selection is compared to three conditions: FDS classification without selecting a subset of dissimilarity features, FDS classification with a randomly selected subset of dissimilarity features and a linear SVM classifier trained by using the original feature vector data. The results again suggest that compactness based selection is beneficial in the context of highly complex data. Two commonly known datasets were used in this experiment, namely the Iris and the Bupa data. The Iris data is a well known dataset in which the class labels are easily separated by a linear decision boundary, this is similar to the linear data as presented in Figure 2.2C. In both these datasets the use of FDS classification seems less beneficial. The Bupa dataset is used in several papers and the class labels are best separated by a flexible decision boundary (Jiang and Zhou, 2004). The results of the experiment in section 2.6 indicate that compactness based selection is only beneficial in combination with the Bupa data. This indicates that compactness based selection is beneficial when the data is optimally separated by a flexible decision boundary.

The advantage of compactness based selection is that the method does not depend on a specific classification model and is easily applied over an entire range of classifiers. Additionally, compactness based selection is easily generalised to continuous features in the original feature vector space. However, the performance of compactness based selection in the original feature vector space has not been evaluated yet. Noteworthy is that the choice of a cutoff value does not only depend on a 10-fold cross validation procedure. The cutoff value could be user defined or other methods could easily by applied to find an optimal cutoff value. All these aspects of compactness based selection makes it a potential and dynamic feature selection method in as well the dissimilarity space and the feature vector space.

Overall the results in section 5 indicate that in terms of the performs FDS classification is not always beneficial. The model with the lowest misclassification rate and the highest AUC is the random forests model while using the original feature vector data. The random forests model with the original feature vector data performs on average 2% better than FDS classification applied in the context of a random forests model. These results are hardly comparable to other classification techniques since a random forests classifier is characterised by a decision boundary that is formulated by an ensemble of classification trees which uses a rectangular decision boundary to separate the class labels. At least in this example we could easily argue that the use of the original feature vector data is preferred over FDS classification. However these results are not to be generalised to other datasets. Further research should be conducted to investigate in which conditions FDS classification is beneficial in the context of a random forests model.

In section 5 linear support vector machines (SVM) in combination with the original feature vector data is compared to FDS classification in combination with a linear SVM. FDS classification clearly outperforms the linear SVM in terms of the misclassification rate and the AUC. The performance of FDS classification is comparable to the performance of radial base SVM. Both techniques obtained a misclassification rate of 0.17. Only the AUC differs slightly. The almost identical results are easily explained by the fact that both classifiers are characterised by a flexible decision boundary. It becomes interesting if we compare the results of radial base SVM with the results of FDS classification applied in the context of logistic regression or a linear SVM. The performance of these two techniques is almost identical in terms of the misclassification rate and the AUC. Both techniques uses the Euclidean distance function, however, radial base SVM uses a similarity measure that is bounded by one and FDS classification uses a dissimilarity measure. A similarity measure indicates that if two observations are identical their pairwise similarity measure is equal to one. Lets assume a pairwise dissimilarity between object i and r is formulated by using the Euclidean distance and is denoted as d_{ir}. In the case of radial base SVM the pairwise similarity is formulated by using $\exp(-d_{ir}^2)$. Despite that these two classification methods differ in how they use the Euclidean distance they seem to achieve the same results in terms of the misclassification rate and AUC. We could not find a legitimate explanation why the results of these two classifiers are almost identical.

The results of FDS classification in the context of a linear SVM suggest a similar trend as compared to FDS classification in combination with logistic regression. Both techniques seem to achieve almost identical misclassification rates. However, the difference between logistic regression and linear SVM is minimal since both techniques are characterised by a linear decision boundary. However, logistic regression aims to find a linear decision boundary that separates the class labels while a linear SVM finds a decision boundary that separates the class labels with a maximum margin around the decision boundary. The similarities between linear SVM and logistic regression have been studied and the results indicate that the performance in terms of misclassification rate is almost identical (Hastie et al., 2009).

During the process of the bootstrap simulation, as described in the methodology section, several obstacles were encountered due to combination of FDS classification and the principal of bootstrapping. Bootstrapping is characterised by generating a new sample by sampling with replacement from a given dataset. Due to principal of sampling with replacement it is likely that a newly generated sample of size n accommodates identical observations in the feature vector space. In the case of FDS classification the newly generated sample in the original feature vector space is transformed into a dissimilarity matrix of size $n \times n$. Now each observation is also represented as a dissimilarity feature in the dissimilarity matrix. As a consequence of sampling with replacement it is likely that identical dissimilarity features are observed in the dissimilarity matrix. In the case of identical features the formulation of a quadratic discriminant analysis (QDA) classifier becomes impossible since it is no longer possible to invert the covariance matrix.

Identical features are also problematic in the case of Naive Bayes which makes a strong assumption about the independence of the features. In these two methods the combination of FDS classification and bootstrapping is not optimal, and in the case of QDA even impossible. However, such a problem could also occur while using original data, for example lets assume a data set with three categorical features with each two labels and a two class outcome variable. When the number of observations increases it becomes more likely that some observations are identical. To overcome this problem one could repeat a cross validation procedure to get an estimate of the misclassification rate and the AUC. An alternative is achieved by repeatedly generating a random training and validation set from the original data. However, in this case the samples are not independently sampled. This procedure is commonly known as the holdout method.

The disadvantage of FDS classification is that the coefficients provided by several models such as logistic regression do no longer provide any information about the value of the original features. In the case of logistic regression in combination with FDS classification the estimated coefficients represent the effect of a specific dissimilarity feature on the outcome variable. As a consequence we only obtain a set of regression coefficients for a set of dissimilarity features towards individuals in the training set. Nonetheless, during the construction of a classifier we are often interested in the relation between a specific feature and the outcome variable.

In the case of logistic regression the importance of an original feature is identified by a permutation approach. In this permutation approach the values of a specific feature are permuted before transforming it into a dissimilarity structure. The importance measure in the context of logistic regression is the deviance obtained by using the original data minus the deviance while using the permuted data (De Rooij, 2015). By repeating this permutation procedure an estimate of the importance per feature is obtained (Ho, 1998). This concept of measuring importance could be generalised to different models. Instead of using the deviance as a measure of importance, the average decrease in accuracy or a measure of goodness of fit are possible candidates for evaluating the importance per feature.

A second alternative and model independent method is the use of the compactness measure of a dissimilarity matrix. This compactness measure is defined in expression 2.4 and is the average of all the compactness measures for each individual dissimilarity feature. In this newly proposed method we also use a permutation approach. Again the values of a specific feature are permuted before transforming it into a dissimilarity structure. After transforming the data with the permuted feature into a dissimilarity matrix an estimate of the compactness is formulated by using expression 2.4. The importance measure is estimated as the compactness measure while using the original data minus the compactness measure while using the permuted data.

The results of the model independent importance measure are discussed in section 5. However, this newly proposed technique has not been evaluated. The results of this importance measure indicate a similar trend for the top three most important features as the importance measure that is incorporated in the random forests classifier in combination with the original feature vector data. However, for the remaining features different results are obtained. To assess validity of this newly proposed method future research should be conducted to evaluate if this technique is a legitimate estimator of the importance for individual features.

6.2 Conclusion Clinical Implications

In the practical problem the aim was to correctly identify individuals with a persistent or recurrent depression by using a set of bio-makers and a short disability questionnaire. In some sense this is the holy grail for psychiatrists since psychiatric diagnostics are primarily derived by a clinical interview. This is in contrast to other medical conditions such as cancer, hiv/aids, diabetes and infectious diseases which are often diagnosed by using blood samples. The advantage of using biological markers in diagnosing individuals with a depressive disorder is that it is less prone to human error. In the case of diagnosing depression it has been shown that the accuracy of diagnosing a depression correctly decreases if individuals have psychiatric co-morbidity (Nuyen et al., 2005).

Noteworthy is that the label of a persistent or recurrent depression has been derived by using a standardised psychiatric interview (CIDI). Research has shown that the CIDI interview has a high false positive rate which falsely elevates the prevalence of a depression (Kurdyak and Gnam, 2005). Additionally the CIDI interview has been criticised for its insensitivity to cultural differences (Rosenman, 2012). This indicates that the outcome variable as used in this thesis may contain false positives which may influence the misclassification rate and the AUC of the learning algorithms. An alternative is use unsupervised learning algorithms to detect structural differences in biological markers that potentially could serve as an objective marker to diagnose a persistent or recurrent depression.

A random forests classifier in combination with the original feature vector data is the superior model in terms of the misclassification rate and the AUC. Due to the nature of a random forests model the importance of each individual feature in terms of the average decrease in accuracy can be estimated. The average decrease in accuracy is interpreted as the average decrease in accuracy of a classification tree when a specific feature is permuted. The results as discussed in section 5 indicate a superior predictive value for creatinine with respect to the remaining features. The average decrease in accuracy after removing creatinine is 12.8%. However, the exact mechanism of creatinine on the prevalence of a persistent or recurrent depression is not exactly clear and a limited amount scientific literature is available about this topic. In section 3, the description of the data reveals that the average serum expression of creatinine is 90.4 for individuals diagnosed with a persistent or recurrent depression and 76.04 for individuals without a persistent or recurrent depression. Research suggest that the severity of a depression is related to the creatinine serum expression, higher levels of creatinine are related with more severe depressive symptoms (Segal et al., 2007). One of the arguments is that the functional neurotransmission is dependent on the intracellular energy metabolism which is partly supported by the expression of creatinine (Allen, 2012).

The random forests model also identified two other important features, namely the expression of glucose and aspartate transaminase (ASAT). The relation between glucose has been well studied although the exact biological relation is still under debate. Some argue that depression is a risk factor for type 2 diabetes while others suggest that type 2 diabetes is a risk factor for depression (Weber et al., 2000). However there is no scientific literature available that studies the mechanism between the expression of ASAT and depression. The data suggest a decrease in the expression of ASAT in individuals diagnosed with a persistent or recurrent depression. However, previous research that investigated the relation between the expression of ASAT and depression revealed an increase in the serum levels of ASAT in individuals diagnosed with a major depression (Zelber-Sagi et al., 2013). These results are in contrast to the findings presented in this thesis. A possible mechanism that could explain the lowered levels of ASAT is due to the fact that some individuals within the NESDA study use anti-depressive medicine which may alter the expression of ASAT. Further research is needed to investigate why we observed a decreased expression of ASAT in individuals with a persistent or recurrent depression.

Noteworthy is that the results indicate that the features are not capable of predicting the presence of a persistent or recurrent depression. If individuals are diagnosed with a single depressive disorder at baseline and diagnosed with a persistent or recurrent depression three years after baseline the blood samples does not contain a pattern that may indicate the presence of a persistent or recurrent depression after three years. These results indicate that changes in the blood serum samples are mainly due to the presence of a persistent or recurrent depression.

The results in section 5 indicate that by using a combination of bio-markers and a disability scale, learning algorithms such as logistic regression, support vector machine and linear discriminant analysis are reasonable well competent in discriminating between individuals with and without a persistent or recurrent depression. The overall performance in terms of the misclassification rate and AUC is possibly improved if other bio-markers are added into the set of features. Examples are the use of genetics or the formulation of a genetic risk score which could possibly improve the overall performance of the classification models. The overall performance could also be improved by increasing the training set size, during the process of this thesis we used a training set of size $n = 500$ by increasing it to $n = 1000$ the performance of the classifiers in terms of the misclassification rate and the AUC are improved. However, given the results as discussed in section 5 it is reasonable to assume that by improving the performance of the learning algorithms psychiatrist could use a blood sample to get an indication of how likely it is that an individual has a persistent or recurrent depression. This indication, as given by a learning algorithm, could be used in combination with a clinical interview to decrease the overall false positive rate.

7 Discussion

In section two compactness based selection is presented as a newly proposed method for selecting a subset of dissimilarity features. This method is derived from the compactness hypothesis (Arkadev and Braverman, 1966) which gives the fundamentals for estimating the complexity of a classification task. In this thesis we generalised this concept to individual dissimilarity features which basically estimates the compactness for each individual dissimilarity feature. By using the compactness measure for each individual dissimilarity feature its possible to select a subset of dissimilarity features.

Assume a compactness estimate for a single dissimilarity feature of 0.5. This indicates that on average 50% of the pairwise dissimilarities towards observations with the same class label as the observation associated with the dissimilarity feature is smaller than the pairwise dissimilarities towards observations with a different class label. Thus, a compactness estimate of 0.5 indicates that the specific dissimilarity feature is barely informative for a classification model and only contains noise. In the case of a compactness estimate of one unit the class labels are perfectly separated by the pairwise dissimilarities. In this thesis a 10-fold cross validation procedure was applied to find an optimal cutoff value between 0.5 and the maximum compactness estimate for a specific dataset. This concept is displayed in Figure 7.1A in which the density of the compactness measures is displayed for the complete data as used in section 5. This figure reveals that by finding the optimal cutoff value for compactness based selection between 0.5 and the maximum compactness measure we assume that compactness measures below 0.5 are not informative for a classification model. However, a compactness measure below 0.5 is likely to be just as informative as a dissimilarity feature with a compactness measure above 0.5. For example, a compactness measure of 0.25 indicates that the observation associated with the specific dissimilarity feature is on average closer to observations with a different class label. These observations associated with a dissimilarity feature that has a compactness measure below 0.5 could for example represent an outlier but still be informative while classifying.

In Figure 7.1B an alternative to the currently used method for compactness based selection is presented. In this newly proposed method two optimal cutoff values are found by a 10-fold cross validation procedure. The extra optimal cutoff value is found between 0.5 and the minimum compactness measure of a specific dataset. Due to this adjustment dissimilarity features with a compactness measure below 0.5 are now considered to be informative and used while training a classifier. As a consequence of this adjustment it is likely that FDS classification models contain more information and thereby achieve better performance.

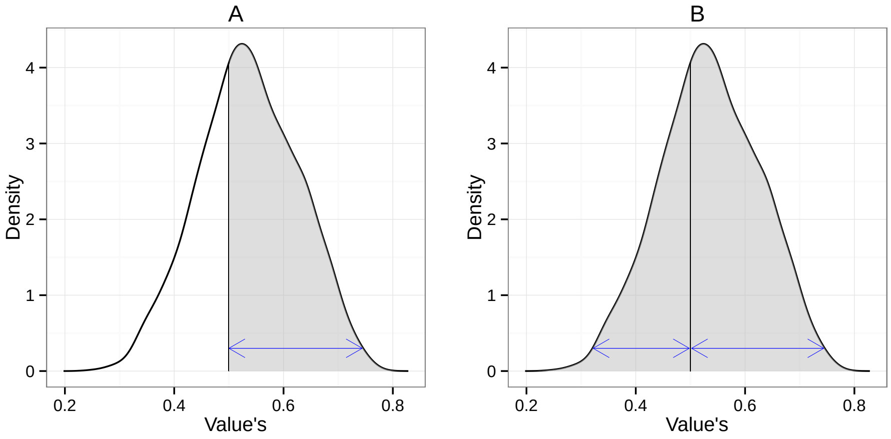


Figure 7.1: Density curve of the compactness measures. In Figure A the optimal cut-off value for compactness based selection is found in the grey area under the curve, between 0.5 and the maximum compactness measure. Figure B represents a newly proposed method in which a second cut-off value is found below 0.5.

A disadvantage of FDS classification is the incapability of providing any information about the importance of individual features in the original feature vector space. The importance measure as described in section 4 makes it possible to assess the additional value of a specific feature on the outcome variable. An alternative is to formulate a completely new classification model that optimises a set of coefficients for features in the original feature vector space before transforming it into dissimilarity structure. In the current framework the Euclidean distance between object 1 and 2 with representations on p features is defined as:

$$d_{12} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \ldots + (x_{1p} - x_{2p})^2} \tag{7.1}$$

All the dissimilarities within a specific feature are weighted by a coefficient in this new classification algorithm. These coefficients are optimised by a yet to be developed optimisation algorithm in such a fashion that the advantages of FDS classification are preserved. In the case of this new model the distance between between object 1 and 2 with values on p features is defined as:

$$d_{12} = \sqrt{\beta_1(x_{11} - x_{21})^2 + \beta_2(x_{12} - x_{22})^2 + \ldots + \beta_p(x_{1p} - x_{2p})^2} \tag{7.2}$$

In order to preserve the properties of FDS classification a new optimisation procedure that optimises the β coefficients has to be developed. The potential of this model is that the dissimilarities between specific features are weighted accordingly to their importance. This allows the user to select features in the original feature vector space while classifying in the dissimilarity space.

The results of this thesis indicate that FDS classification is beneficial in the context of linear classifiers in which FDS classification finds a non-linear decision boundary. These non-linear boundaries fitted by FDS classification are a good competitor for radial base SVM's. The advantage of FDS classification over SVMs is that the output is interpretable in terms of dissimilarities towards prototypes or exemplars. SVM's are often characterised as black boxes in which it is unknown what is done in order the classify. However, a large extend of research is needed in the field of formulating prototypes. Most common techniques only allow the input of continuous features, methods for defining prototypes with categorical features are often insufficiently explored.

In the case of FDS classification in the context of logistic regression the use of compactness based selection or L_1 regularisation does provide a superior tool for selecting a subset of dissimilarity features while the classification task is considered to be complex. For less complex classification tasks forward stepwise selection is a good competitor of compactness based selection and L_1 regularisation. Using FDS classification in the context of logistic regression without selecting a subset of dissimilarity features is proven unstable since the coefficients are not uniquely estimated.

In this thesis we mainly used the Euclidean distance. The complexity of a classification task as measured by the compactness hypothesis is dependent on the distance function used to transform the original feature vector data into dissimilarity data. A large collection of different distance functions are available and could improve the performance. The Mahalanobis distance is good competitor for the Euclidean distance since it takes into account the variances and covariances between features in the original feature vector space. The perfect distance function could differ per classification task, although it is likely that some of the distance functions in general are more suitable for FDS classification. Estimating the complexity of a classification task requires a metric distance function. The use of asymmetric distance function in combination with FDS classification is hardly studied and new methods for estimating the complexity and selecting dissimilarity features are required in that case.

In this thesis we standardised all the original features before transforming the original feature vector data into a dissimilarity structure. However, the standardisation of the original features is not a necessity. However, if the original features are not standardised the importance of the original features are a function of the scale of the original features. Multiple alternatives are available, the original features could be used without standardising and instead features could be weighted accordingly to their importance.

Overall the results indicate that FDS classification is beneficial in combination with a linear classifier and a complex classification task. However, further research is needed to identify when FDS classification is beneficial in a non-linear classifier such as radial base SVM or a random forests classifier.

All the experiments are conducted by using R (R Core Team, 2014) and the R code for all the functions are available on request by sending an e-mail to n.jongs@me.com. The data that accommodates all the biological markers which are related to psychiatric disorder is not publicly available. The data is available on request by contacting the NESDA consortium (Penninx et al., 2008).

References

Acheson, A., Conover, J., Fandl, J., De Chiara, T., Russell, M., Thadani, A., Squinto, S., Yancopoulos, G., and Lindsay, R. (1994). A bdnf autocrine loop in adult sensory neutrons prevents cell death. *Nature*, 374:450–453.

Allen, P. (2012). Creatine metabolism and psychiatric disorders: Does creatine supplementation have therapeutic value? *Neuroscience and Biobehavioral Reviews*, 36(5):1442–1462.

Anderson, R., Freedland, K., Clouse, R., and Lustman, P. (2001). The prevalence of comorbid depression in adults with diabetes. *Diabetes Care*, 24(6):1069–1078.

Annwiler, C., Allali, G., Allain, P., Bridenbaugh, S.and Schott, A., Kressig, R., and Beauchet, O. (2009). Vitamin d and cognitive performance in adults: a systematic review. *Eur J Neurol*, 16:1083–1089.

Arkadev, A. G. and Braverman, E. M. (1966). *Computers and Pattern Recognition.* Washington, D.C., Thompson.

Arzheava, Y., Tax, D. M. J., and van Ginneken, B. (2009). Dissimilarity-based classification in the absence of local ground truth: Application to the diagnostic interpretation of chest radiographs. *Pattern Recognition*, 42:1768–1776.

Ashby, G. and Maddox, T. (1993). Relations between prototypes, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37:372–400.

Berk, M., Williams, L., Jacka, F., O'neil, A., Pasco, J., Moylan, S., Allen, N., Stuart, A., Hayley, A., Byrne, M., and Maes, M. (2013). So depression is an inflammatory disease, but where does the inflammation come from ?. *BMC Medicine*, 11(1):11–200.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cherniack, E., Troen, B., Florez, H., and Levis, S. (2009). Some new food for thought: the role of vitamin d in the mental health of older adults. *Curr Psychiatry Rep*, 11:12–19.

De Rooij, M. (2015). The delta-machine. Not Published.

Dowlati, Y., Herrmann, N., Swardfager, W., Liu, H., Sham, L., Reim, E. K., and Lanctôt, K. L. (2010). A meta-analysis of cytokines in major depression. *Biological Psychiatry*, 67(5):446–457. Cortical Inhibitory Deficits in Depression.

Duin, R. P. W. (1999). Compactness and complexity of pattern recognition problems. In *International Symposium on pattern Recognition 'In Memoriam Pierre Devijver'*, pages 124–128. Royal Military Academy, Brussels.

Duin, R. P. W., Loog, M., Pekalska, E., and Tax, D. (2010). Feature-based dissimilarity space classification. *Lecture Notes in Computer Science*, 6388:46–55.

Duin, R. P. W. and Pekalska, E. (2006). Dissimilarity-based classification for vectorial representations. *Pattern Recognition*, 3:137–140.

Duin, R. P. W. and Pekalska, E. (2010). Non-euclidean dissimilarities: Causes and informativeness. In Hancock, E. R., Wilson, R. C., Windeatt, T., Ulusoy, I., and Escolano, F., editors, *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR&SPR 2010, Cesme, Izmir, Turkey, August 18-20, 2010. Proceedings*, volume 6218 of *Lecture Notes in Computer Science*, pages 324–333. Springer, Springer.

Duin, R. P. W. and Pekalska, E. (2012). The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826–832.

Edelman, S. (1999). *Representation and Recognition in Vision.* MIT Press, Cambridge.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99:619–632.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York, NY.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fisher, R. A. (1936). The use of mutiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Goldfarb, L. (1985). A new approach to pattern recognition. In Kanal, L. N. and Rosenfeld, A., editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers BV.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning:Data Mining, Inference, and Prediction.* Springer, 2 edition.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *World Scientific*, pages 21–34.

Jiang, Y. and Zhou, Z. (2004). Editing training data for knn classifiers with neural network ensemble. In *Lecture Notes in Computer Science, Vol.3173*, pages 356–361. Springer.

Kessler, R. C., Abelson, J., Demler, O., Escobar, J. I., Gibbon, M., Guyer, M. E., Howes, M. J., Jin, R., Vega, W. A., Walters, E. E., Wang, P., Zaslavsky, A., and Zheng, H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *Int J Methods Psychiatr Res*, 13(2):122–139.

Klein, S., Loog, M., Lijn, F., Heijer, T., Hammers, A., Bruijne, M., Lugt, A., Duin, R., Breteler, M., and Niessen, W. (2010). *Early diagnosis of dementia based on intersubject whole-brain dissimilarities*, pages 249–252. IEEE Press.

Kurdyak, P. A. and Gnam, W. H. (2005). Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry*, 50(13):851–856.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.6-6.

Milaneschi, Y., Hoogendijk, W., Lips, P., Heijboer, A., Schoevers, R., Van Hemert, A., Beekman, A., smit, J., and Penninx, B. (2014). The association between low vitamin d and depressive disorders. *Molecular Psychiatry*, 19:444–451.

Neveu, I., Naveilhan, P., Jehan, F., Baudet, C., Wion, D., Luca, H. F. D., and Brachet, P. (1994). 1,25-dihydroxyvitamin d3 regulates the synthesis of nerve growth factor in primary cultures of glial cells. *Molecular Brain Research*, 24(1):70–76.

Nuyen, J., Volkers, A. C., Verhaak, P. F., Schellevis, F. G., Groenewegen, P. P., and Van den Bos, G. A. (2005). Accuracy of diagnosing depression in primary care: the impact of chronic somatic and psychiatric co-morbidity. *Psychol Med*, 35(8):1185–1195.

Pekalska, E., Duin, R. P. W., and Paclík, P. (2006). Prototype selection for dissimilarity-based classifiers. *Pattern Recogn.*, 39(2):189–208.

Pekalska, E., Paclik, P., and Duin, R. P. W. (2001). A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211.

Penninx, B., Beekman, A., Smit, J., Zitman, F., Nolen, W., Spinhoven, P., Cuijpers, P., De Jong, P., Van Marwijk, H., Assendelft, W., Van Der Meer, K., Verhaak, P., Wensing, M., De Graaf, R., Hoogendijk, W., Ormel, J., and Van Dyck, R. (2008). The netherlands study of depression and anxiety (nesda): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17:121–140.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rosenman, S. (2012). Cause for caution: culture, sensitivity and the World Mental Health Survey Initiative. *Australas Psychiatry*, 20(1):14–19.

Segal, M., Avital, A., Drobot, M., Lukanin, A., Derevenski, A., Sandbank, S., and Weizman, A. (2007). Serum creatine kinase level in unmedicated nonpsychotic, psychotic, bipolar and schizoaffective depressed patients. *European Neuropsychopharmacology*, 17(3):194–198.

Sharma, R., Tun, N., and Grayson, D. (2008). Depolarization induces downregulation of dnmt1 and dnmt3 in primary cortical cultures. *Epigenetics*, 3(2):74–80.

Tanskanen, A., Tuomilehto, J., and Viinamäki, H. (2000). Cholesterol, depression and suicide. *The British Journal of Psychiatry*, 176(4):398–399.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 4(84):327–352.

Ulas, A., Duin, R. P. W., Castellani, U., Loog, M., Mirtuono, P., Bicego, M., Murino, V., Bellani, M., Cerruti, S., Tansella, M., and Brambilla, P. (2011). Dissimilarity-based detection of schizophrenia. *Int. J. Imaging Systems and Technology*, 21(2):179–192.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory.* Springer, New York.

Warner-Schmidt, J. and Duman, R. (2006). Hippocampal neurogenesis: opposing effects of stress and antidepressant treatment. *Hippocampus*, 16(3):239–249.

Weber, B., Schweiger, U., Deuschle, M., and Heuser, I. (2000). Major depression and impaired glucose tolerance. *Exp. Clin. Endocrinol. Diabetes*, 108(3):187–190.

Yamada, K. and Nabeshima, T. (2003). Brain-derived neurotrophic factor/trkb signalling in memory processes. *Journal of Pharmacological sciences*, 91(4):267–270.

Zelber-Sagi, S., Toker, S., Armon, G., Melamed, S., Berliner, S., Shapira, I., Halpern, Z., Santo, E., and Shibolet, O. (2013). Elevated alanine aminotransferase independently predicts new onset of depression in employees undergoing health screening examinations. *Psychol Med*, 43(12):2603–2613.