DEGREE PROJECT IN COMPUTER ENGINEERING,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2018

Evaluation of Feature Selection Methods for Machine Learning Classification of Breast Cancer

THONY PRICE

NIKLAS LINDQVIST

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Evaluation of Feature Selection Methods for Machine Learning Classification of Breast Cancer

NIKLAS LINDQVIST
THONY PRICE

Degree Project in Computer Science
Date: June 4, 2018
Supervisor: Pawel Herman
Examiner: Örjan Ekeberg
Swedish title: Evaluering av Attributurvalsmetoder för klassificering av bröstcancer med maskininlärning
School of Electrical Engineering and Computer Science

Abstract

Breast cancer is the leading cause of cancer deaths among women today. Computer aided diagnosis has proved efficient in assisting medical experts to set an early diagnosis improving the chance of recovery. Computer aided diagnostics utilizes machine learning to make a prediction whether a patient has a benign or malignant cancer. For this purpose, machine learning algorithms are used to perform classification. Applying feature selection the algorithms can be fed data with lower dimensionality and can produce a more accurate result. In this report we conducted experiments with four different feature selection methods and four classifiers on four datasets.

We found that Artificial neural networks have a significant increase in classification accuracy of breast cancer when applying feature selection. The maximum improvement in accuracy was 51% using the feature selection method Entropy and data from Royal Hallamshire Hospital. The accuracy achieved by artificial neural networks does not show any correlation with a specific feature selection method. Using Naïve Bayes, Support Vector Machines and Decision trees no increase in accuracy using feature selection could be statistically proven considering all datasets. However, in some observations these classifiers manifested increased classification accuracy with feature selection compared to using all features of the dataset.

Sammanfattning

Bröstcancer är idag den cancerform som orsakar flest dödfall hos kvinnor. Datordriven diagnostisering har visat sig effektiv i att assistera medicinska experter med att sätta en tidig diagnos för cancer och därmed öka chanserna för tillfrisknande hos patienten. Datordriven diagnostisering använder sig av maskininlärningsmetoder för att göra en prediktion huruvida en tumör är god- eller elakartad. I denna diagnosticeringsprocesses används en patients data av en maskininlärningsalgoritm för att göra automatisk klassificering. Applicerandet av attributurvalsmetoder innebar att algoritmen kan använda sig av data med färre dimensioner och producera ett mer träffsäkert resultat. Vi genomförde experiment med fyra attributurvalsmetoder, fyra maskininlärningsalgoritmer och fyra dataset.

Vi fann att artificiella neurala nätverk med hjälp av attributurvalsmetoder visar en signifikant ökning av träffsäkerhet vid klasificering av bröstcancer. Den maximala förbättringen var 51% då attributurvalsmetoden entropi användes i kombination med data från Royal Hallamshire Hospital. För artificiella neurala nätverk kunde vi inte finna något samband mellan vilken attributurvalsmetod som användes och uppnådd träffsäkerhet, detta varierade från fall till fall. För metoderna Naïve Bayes, Support vector machine och Beslutsträd kunde ingen signifikant ökning av träffsäkerhet fastställas vid användning av attributurvalsmetoder. Dock kunde i vissa fall en ökning av klassificeringsträffsäkerhet observeras med hjälp av dessa metoder jämfört med klassificering med alla attribut.

Contents

Chapter 1

Introduction

Breast cancer is a disease of major concern and is the leading cause of cancer deaths among women [3]. At present there are no effective ways to prevent breast cancer. However, efficient diagnosis in an early stage can increase the chance of full recovery. This makes early detection and diagnosis an important issue where currently mammography screenings is the primary imaging modality for early detection of breast cancer [26].

Hospitals today collect data to do monitoring and some of this data, such as mammography screenings can be collected and shared in information systems. The data can be used by medical personnel to increase their understanding of different diseases. It can also be used in computer aided diagnostics (CAD) where machine learning algorithms enable tools for intelligent data analysis. CAD makes use of machine learning techniques that learn a hypothesis, a statistical prediction about a patient's diagnosis from a large set of previously diagnosed examples. The overarching purpose is to assist medical experts in more efficient and accurate diagnostics [18]. Machine learning algorithms is a well studied field within medical diagnosis and well suited for analyzing medical data, especially within small specialized diagnostic problems such as breast cancer [15].

Multiple studies of CAD on breast cancer have been conducted, primarily focusing on classifying mammography data of tumors as malignant or not, such those of Ramos-Pollán et al. [23] and Akay [2]. The act of feature selection, removing redundant or irrelevant features

from a dataset, can provide classifiers to be faster, more cost-effective and accurate. With feature selection the understandability can be improved which is a clear benefit when it comes to medical decisions [17]. It is also explicitly mentioned as a topic in need of more research in studies made on breast cancer diagnostics [20].

1.1 Research Question

In our thesis we will study the impact of four feature selection methods on the classification rate of malignant breast cancer by four different machine learning methods. We aim to answer the following:

- Does the feature selection improve the accuracy of classification compared to using all features?
- Is the effect of feature selection dependent on the classification method?

Our hypothesis is that overall the feature selection will improve the classification rate of the machine learning methods used in the context of breast cancer classification. This hypothesis is based on previous research reported by Karabulut, Özel, and İbrikçi [17], where it was found that classification accuracy on 15 different datasets of medical and non-medical data was increased by the use of filtering methods for feature selection.

Our research differs from the work presented in [17] in the amount of datasets used and number of classification methods. In our project we will put more emphasis on breast cancer using only datasets of that type. The previous work only investigated feature selection methods by filtering which we will extend by implementing wrapper methods. Lastly, the research scope in this thesis includes a study of the effect of different feature selection methods on Support Vector Machines (SVMs), which was not included in [17].

1.2 Approach

Trials will be conducted with feature selection by using both wrapper methods and feature selection filter methods. The result of the feature selection methods will be used with different classifiers to evaluate their performance. To broaden the base for comparison we will use several classifiers, Decision Tree (DT) a logic based algorithms, Artificial Neural Network (ANN) a perceptron-based technique, Naïve Bayes (NB) a statistical learning algorithms and Support Vector Machines (SVM) [27]. These classifiers are commonly used in CAD and thus relevant to study [23], [2], [18]. Feature selection (FS) methods and classifiers included in this report are denoted in table 1.1.

Included classifiers and FS-methods

Classifiers	Artificial Neural Network (ANN) Decision Tree (DT) Naïve Bayes (NB) Support Vector Machine (SVM)
FS by Wrapping	Sequential Backward FS (SBS) Sequential Forward FS (SFS)
FS by Filtering	Chi-square (Chi2) Entropy

Table 1.1: All classifiers and feature selection methods included in this paper. Each classifier will be tested with each FS-method yielding 16 distinct combinations.

A comparison between the classification rate of the machine learning methods without using any feature selection, and classification rate when using feature selection will be conducted. The comparison may then establish the importance of feature selection in different machine learning approaches when classifying breast cancer. The evaluation of impact by FS will be measured by computing the ratio between best accuracy achieved with and without using feature selection. Several datasets with different types of features will be used for evaluation. Details on the datasets will be presented in section 3.1. The reason for using multiple datasets is to strengthen the basis for statistical evaluation and possibility conclude a more generalized result.

1.3 Scope

The scope of this thesis is limited by the amount of datasets, classifiers and FS-methods included. Having four of each, the aim is to conclude a result that generalize well. However, it must be considered these are a small selection of all the available possibilities. The data represents a few of the possible ways to collect data when making breast cancer diagnostics. There are many more classifiers and each can be tuned into countless of configurations. Regarding the feature selection methods we evaluate two filter methods and two wrapper methods, it exists many more and also embedded methods which is not included in out thesis.

These limitations results in constraining our scope to the classifiers and FS-methods presented in table 1.1 and the datasets described in 3.1.

Chapter 2

Background

2.1 Breast Cancer

A study in Sweden by Tabár et al. [26] found breast carcinoma mortality was reduced by 63% after mammography was introduced. This clearly emphasize the benefits of screening which has resulted in an increased usage of this method to detect and diagnose breast cancer. The increasing demand for mammography image interpretation has lead to a shortage of medical radiologist to perform this task, consequently non medical personnel supplement the mammography image interpretation [9]. As breast cancer still continues to be the leading cause of cancer mortality among women and more efficient diagnostics and pathology is high on demand, the need of low-cost point-of-care is very large [19].

Fine needle aspiration (FNA) is a diagnostic tool to aspirate cell samples by sampling cells from a tumour, then staining them, and examine under a microscope [28]. An example of such sample can be seen in figure 2.1. The cell samples can be evaluated within 24 hours and the method is cost-effective and can be used as a preoperative tool for investigation of tumours. The method is also complication-free and has been widely used for the past 60 years.

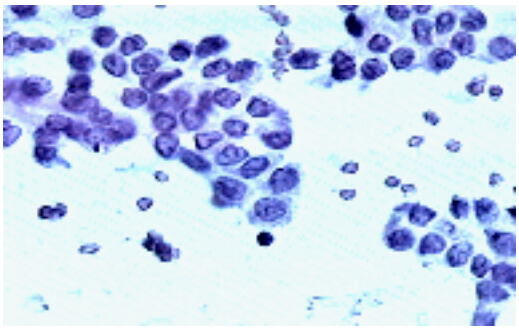


Figure 2.1: A caption of a FNA sample as seen through a microscope. The sample is collected from a breast lump by fine needle aspiration then stained to engender the features of the cells. Image courtesy of Dheeru and Karra Taniskidou [11].

2.2 Computer Aided Diagnostics

Machine learning techniques have been successfully applied to computer aided diagnostics (CAD). By a computerized procedure it provides a second, objective opinion of medical image interpretation and diagnosis [18], [8]. To create a CAD system, samples with a diagnose is firstly gathered and stored, then used for learning. In the case of breast cancer detection, a radiologist put labels on a set of mammography scans. These include the diagnosis of the scan and possibly additional attributes connected to the scan too, such as patients age or other conditions. These scans together with the labels can then be used to learn a hypothesis whether a undiagnosed sample contains benign or malignant cancer [18].

2.3 Machine Learning

The field of machine learning is concerned with automated discoveries of regularities in data with use of computer algorithms. These regularities can then be used to take actions, such as classifying data into different categories or making predictions [6]. As the data may differ from images to population growth to medical data, the computer algorithms differ too. The algorithms used in this report is detailed below.

2.3.1 Artificial Neural Networks

The term Artificial Neural Network (ANN) originates from the structure of the algorithm. The structure vaguely mimics the biological network of a human brain [6]. A representation of such network is visualized in figure 2.2. The learning process of an ANN is conceptually two parts, feed forward and back propagation.

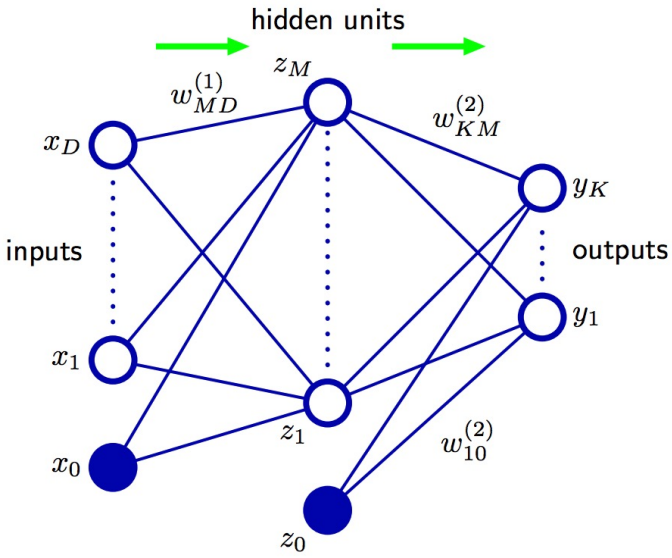


Figure 2.2: Network diagram for a two layer neural network. X_D represents the input data which is linearly transformed by the weights W_{MD} and bias X_0 and forwarded to the hidden units. From the hidden units the data is transformed in the same manner to the output layer y_K which corresponds to the potential classes of the input data. Image courtesy of Bishop [6].

When initially feeding forward the data constitutes the input of the network. The input signal is equal to the dimensionality of the data. Forwarding the signal to the first layer, a linear transformation is applied to the signal involving two parameters conclusive to the layer, weight and bias. The transformed signal is sent through an activation function amplifying or reducing the signal based on its current value, producing a new output. This output is then considered the input for a new set of weights and biases, then forwarding the signal repeatedly in the same manner until reaching the final layer. In the final layer the signal is converted to probabilities, one for each possible output of the network [6].

Back propagation is where an ANN is tuning its wights and biases to produce sensible outputs in relation to the input. Feeding an input of training data forward, an error can be computed from the difference between the received output probabilities and the expected output. The error is in turn propagated backwards, tuning the parameters to produce the expected output instead. Repeating this process for all training data the network can increase its prediction accuracy by tuning itself to the data [6].

2.3.2 Decision Trees

Tree based methods involve stratifying or segmenting the predictor space into a number of simple regions. In order to make a prediction for a given observation, the mean or the mode of the training observations is used in the region to which it belongs. Since the set of splitting rules used to segment the predictor space can be summarised in a tree, these types of approaches are known as Decision Trees (DT) [16].

DTs can be applied both to regression as well as classification problems. We focus on classification DTs as its the nature of the classification at hand in this report.

To grow a classification tree recursive binary splitting is performed. The binary spilt is based on classification error rate on the training data. Since we want the classifier to assign an observation in a given region to the most commonly occurring error rate class of training observations in that region. The classification error rate is simply the fraction of the training observations in that region that do not belong to the most common class [16].

2.3.3 Naïve Bayes

Naïve Bayes (NB) have beens studied since the 1950s and is a supervised, probabilistic machine learning classifier. It uses the posterior which statistically determines the probability of a data point x belonging to a certain class y given the observed data points:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

This formula can be read as:

$$Posterior = \frac{Prior * Likelihood}{Evidence}$$

Where Prior is the prior probability of y belonging to some distribution θ, likelihood the probability of x belonging to distribution θ when having observed y. The evidence is the probability of θ, and finally the Posterior is the conditional probability of y given x. The name of the methods originates from the fact that the maximum posterior is calculated with help of the Bayes' Theorem and the method is naïve in the aspect of it assuming that all features in the data are independent form each other [12].

2.3.4 Support Vector Machines

Support vector machines (SVM) were developed in the 1990s and are based on the simpler maximal margin classifier. The maximal margin classifier creates a hyperplane to separate the data points of different classes from each other. The margin is the smallest perpendicular distance from the plane to a data point. As the name of the method suggests, the goal is to find the plane with the largest margin given the data set [16].

The support vector classifier is an extension to the maximal margin method to try create a method with greater robustness and better classification to most training observations. In order to achieve this goal, a slack variable in introduced to allow points to be on the wrong side of the margin. Lastly, a SVM converts the support vector classifier from a liner separator to a non-linear separator by using so called Kernels. The classifier can now separate more complex data sets with higher accuracy since its not limited to create linear hyperplanes [16].

2.4 Feature Selection

A feature is a variable that describes a data instance. A rectangular surface can be considered having two features, length and height. A rectangular volume can be considered having three features: length,

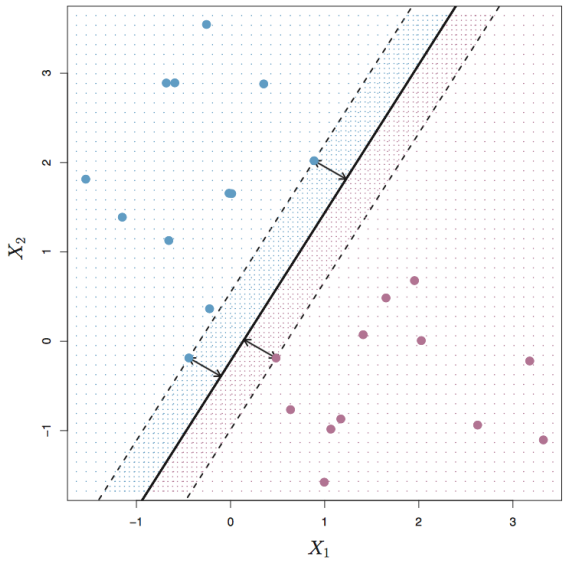


Figure 2.3: Image of SVM margin. The red dots represent one class and the blue dots represent another class. The arrows indicates the maximal margin. Image courtesy of James et al. [16].

height and depth. More complex data instances, such as a gene expression may have up to $60\,000$ features, such vast feature space results in a much harder learning process. Thus, many times it is preferable to select a subset of all available features to reduce the dimensionality [13].

The benefits of selecting a subset of all available features are manyfold, among other it facilitates data visualization and data understanding. It reduces the measurement and storage requirements and reduces training and utilization times. In cases with thousands of features, like the example with gene expressions, it is essential to work with a subset of the data to produce reliable results [13].

2.4.1 Filter Methods

Filter methods are considered a preprocessing step. That means a filter method evaluate features before data is applied to a learning machine, or even before a deciding on a classifier. The evaluation is performed by doing variable ranking by some score, such as information gain.

The score results in a ranking of the attributes and a subset can be selected in order of the ranking [13].

There are both good and bad aspects of filter methods. The positive concerns variable ranking make filter methods very scalable and robust, as the calculations only operates on as many variables as there are features. They can be performed just once and tested on multiple classifiers making it very computational effective. On the other hand, a subset of features that independently might be assessed an non informative, may in combination with other features provide a lot of information to enable good learning [13].

2.4.2 Wrapper Methods

Wrapper methods differs significantly from filter methods. While filter methods are evaluated as a preprocessing step, independent of the classifier, wrapper are evaluated dependent on the classifier.

Wrappers utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power [13]. The issue is that as datasets become large this method might be overly computational intense, as finding the optimal subset is considered to be NP-hard [4].

2.5 Related Work

Medical datasets have been widely used to assess the performance of a multitude of classification strategies and methods. Breast cancer is one of the disease commonly studied within the filed of machine learning [15].

2.5.1 Optimization of Classification Accuracy in CAD

Multiple studies exploring optimal classifiers for CAD have been conducted. Ramos-Pollán et al. [23] tested $20\,000$ classifier configurations to evaluate their ability to correctly classify malignant cancer. They achieved a result of 0.996 under the Receiver operating characteristic

(ROC) curve. The method to achieve this result was a feed forward, back propagating ANN.

With a foundation of well performing classifiers such as [23], studies investigating more fine tuned approaches building upon earlier results have been made. Ozcift and Gulten [20] demonstrated that ensemble learning can be used in CAD to improve the performance of a rotation forest classifier. Using three different dataset and 30 classifying algorithms the average accuracy improved on all datasets by nearly 3%.

Abdel-Ilah and Šahinbegović [1] reported further improvements on ANNs by investigating the optimal number of hidden layers and neurons for a feed forward back propagation network on the WBCD dataset which is included in our experiments as well. The highest accuracy achieved was 98% using 3 hidden layers and 21 neurons with three distinct transfer functions.

2.5.2 Optimization Classification Accuracy in CAD using Feature Selection

There are several ways to optimize a machine learning methodology. Chandrashekar and Sahin [7] explains that feature selection algorithms can be used for a vast amount of benefits such as: simplicity, stability, classification accuracy, storage and computational requirements. They also claim that comparison between different feature selection methods can only be done on one dataset at a time since the underlying algorithms depend on the structure of the data. Improvement in prediction accuracy was achieved in the paper by using feature selection for several datasets to demonstrate the applicability of feature selection techniques.

Akay [2] investigated the performance of classification of a SVM with a RBF kernel using feature selection, filtering by F-score. They achieved a classification accuracy of 99.51% which accordingly was among the highest scores recorded by then (2007). It should be noted, the accuracy was measured on only one dataset meaning the result can not be generalized to datasets at large. The dataset was WBCD.

Babaoglu, Findik, and Ülker [5] demonstrated how binary particle

swarm optimization (BPSO) and genetic algorithm (GA) techniques as feature selection methods improved the accuracy for classifying coronary artery disease using a SVM. The feature selection methods only used 11 features and 12 features respectively compared to the full data set of 23 features and achieved an improvement of 2-4% accuracy.

Karabulut, Özel, and İbrikçi [17] made a comparative study on the effect of feature selection on classification accuracy and found up to 15.55% improvement (an increase from 55.56% to 71.11%) on classification rates using a ANN classifier. The study used only filter algorithms for feature selection, among those were both information and Chi2. The study applied the selected features on three classification methods, Naïve Bayes, Artificial Neural Network as Multilayer Perceptron, and J48 decision tree classifier on 15 different datasets including WBCD.

Building upon this foundation of work our study seeks to complete the field by providing further investigation of unreported selection methods such as SBS and SFS and review their performance on a combination of classifiers from the previous research presented here.

Chapter 3

Method

3.1 Datasets

In this study four different datasets concerning breast cancer was used. An overview of their characteristics is presented in table 3.1 and more detail on their respective content and origin is contained in the following subsections.

Dataset	Size	Features	Ratio (B/M)
Wisconsin (WBCD)	569	30	357/212
Royal Hallamshire Hospital (RHH)	692	11	457/235
Erlangen-Nuremberg (EN)	961	5	516/445
MIAS	119	5	68/51

Table 3.1: All datasets used in the study, their (abbreviation), number of instances, features and ratio between Begin and Malignant samples.

3.1.1 Wisconsin

The Breast Cancer Wisconsin (Diagnostic) dataset, was donated 1995 to UCI Machine Learning Repository [11] by one of its creators, Nick Street. It contains 569 instances with 30 attributes describing the features of breast cancer. Each instance is classified as benign (357) or

malignant (212). The 30 attributes describe ten real-value features of FNA-samples.

3.1.2 Royal Hallamshire Hospital

Fine needle aspirates of breast lumps (FNAB) was collected from 692 patients at Royal Hallamshire Hospital, Sheffield, during 1992 - 1993. The FNABs 10 features of the FNABs was marked as present or non present. These features along with the patients's age defines the attributes of the dataset. In addition, the final outcome of benign disease or malignancy was confirmed by open biopsy where this result was available.

3.1.3 MIAS database

Mias database contain results from 119 data points with 5 features: Character of background tissue, Class of abnormality, X coordinate of centre of abnormality, Y coordinate of centre of abnormality, Approximate radius (in pixels). The features was extracted from 1024x1024 pixel images.

3.1.4 Erlangen-Nuremberg

Dataset collected from a Breast Imaging-Reporting and Data System (BI-RADS) at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. It contains four features assessed as a discrete value from a double-review by physicians along with the patients' age.

3.2 Implementation

Each dataset will be split into training and test data. For each classifier, each FS-method will select all possible number of features in turn. The classifier will be trained on the subset and evaluated on the test data.

A compact pseudocode how results are produced is presented in algorithm 15. The steps is more thoroughly detailed below. Measurements of how results will be evaluated is contained in 3.3.

The methodology is in line with previous research in the field such as [17]. It's because it produces a foundation for comparing the impact of feature selection compared to using the full dataset.

Algorithm 1: Outline of how experiments will be conducted. For each dataset, split into test and training data. Use each classifier in turn to evaluate all feature selection methods on all subset of attributes. For each subset train classifier on training data and evaluate with stratified 10-fold cross validation. Store each result for further analysis.

```
1  for dataset ∈ Datasets do
2      xTrain, yTrain, xTest, yTest = Split(dataset);
3      for classifier ∈ Classifiers do
4          for FS ∈ FSmethods do
5              for num ← 1 to allAtts(xTrain) do
6                  Xtr = selectAttsWithMethod(Xtrain, FS);
7                  Xte = reshapeToAtts(Xtest, Xtr);
8                  clf = train(classifier, Xtr);
9                  fld = stratKFold(10);
10                 res = score(clf, fld, Xte, yTest);
11                 save(res);
12             end
13         end
14     end
15 end
```

3.2.1 Classifiers and Parameters

All classifiers was imported from Scikit [22]. All classifiers allow tuning, by setting parameters of its behaviour. As tuning the parameters for any dataset and/or subset of attributes inflicts bias to the current state all parameters is left to default [10]. Default parameters may cause suboptimal performance of a classifier. However, an optimal performance is not the intention of this experiment, the influence of feature selection is. Thus motivating the default parameter settings.

Those values of the most influential parameters for each classifier can be found in appendix A.1.

3.2.2 Feature Selection

Feature selection with filtering methods was imported from Scikit [22]. The library contains the method "SelectKBest" which transforms data to a subset of k attributes given a method such as Chi2 or Entropy.

Feature selection with wrapper methods was implemented with the "SequentialFeatureSelector" method available in the a library by Raschka [24]. As the wrapper evaluates the performance of each subset when selecting the best, a measurement method of evaluations must be set. The method implemented was classification accuracy as that is what should be used to compare methods at a later stage.

3.3 Evaluation

3.3.1 Test Data and Accuracy

To compare methods and classifiers a measurement is needed. Classification accuracy entails how many labels was correct on a test set, number of correctly classified samples divided by all samples. To ensure fair comparison classification the data split is kept consistent between classifiers by seeding. Test data is only introduced when accuracy is measured to avoid data leakage. These is standard methodology when it comes to evaluating classifiers [16] Test data consist of 30% of dataset. It is maximized by performing stratified 10-fold validation providing 10 accuracy scores on 10 distinct subsets of the test data. The overall performance is computed as the mean over the folds. In this way a more trustworthy estimate of the test accuracy is achieved than if classification were to be evaluated only once on the test set [16].

To evaluate the impact of using feature selection, a measure we will denote *gain* should be computed. It is computed as the ratio between best accuracy using FS, and accuracy not using FS. The accumulated gains of a classifier on all datasets gives a measure of how much it improved

from feature selection, and a basis to compare classifiers against each other.

3.3.2 Differences Among Classifiers and Feature Selection Methods

The method includes four different classifiers and four different FS-methods, these can be considered as separate groups. This results in 16 distinct combinations of classifiers and FS-methods, that is, interaction between these groups. It is needful to evaluate the differences between these groups, and their interaction to fully understand the results.

To evaluate this relationship analysis of variance (ANOVA) test will be performed [25]. ANOVA entails if differences in results between groups can be explained by variance or if there is a statistically significant difference between, or within groups. ANOVA computations of groups and their interactions results in F-scores. F-score measures the probability of rejecting the null hypothesis, that some combination of groups are equal [25].

Chapter 4

Results and Analysis

We performed classification with four different classifiers on four different datasets. In each dataset-classifier combination we measured classification accuracy when applying four different feature selection (FS) methods. The aim is to investigate the impact of feature selection on classification accuracy, and the interaction between different classifiers and FS-methods. To enable a basis for comparisons, classification accuracy was measured without FS too.

4.1 Impact of Factors: Datasets, Classifiers and FS-methods

Before analyzing the results we consider potential interactions between different factors involved in the experiment, these are:

1. Datasets
2. Machine Learning classifiers
3. Feature selection methods

Having four of each factor, there are 16 distinct combinations of classifiers and FS-methods with four measurements each, one on every dataset. The measurement is the best accuracy when applying feature selection, regardless of the size of the subset. All datasets contain different information and therefore introduce necessary variance to the

experiments. The variance allows us to suspect these results generalizes well onto other datasets in the domain of breast cancer, which are not included in this study. In our scope we are particular interested in the interaction of classifiers and FS-methods.

4.1.1 Evaluation of FS-methods and Classifiers

To investigate the effect of feature selection on different classifiers, we performed two-way ANOVA. The values constitutes of four measurements by each of the 16 distinct classifier and FS-method combination. The four measurements come from each of the four datasets. Each measurement is the best accuracy achieved with a attribute subset of the dataset. The ANOVA result is presented in table 4.1.

	RSS	df	F	$P(>F)$	
Classifier	0.3336	3.0	4.660	0.00615	**
Method	0.0099	3.0	0.138	0.93666	
Classifier:Method	0.0631	9.0	0.294	0.97314	
Residual	1.1455	48.0			

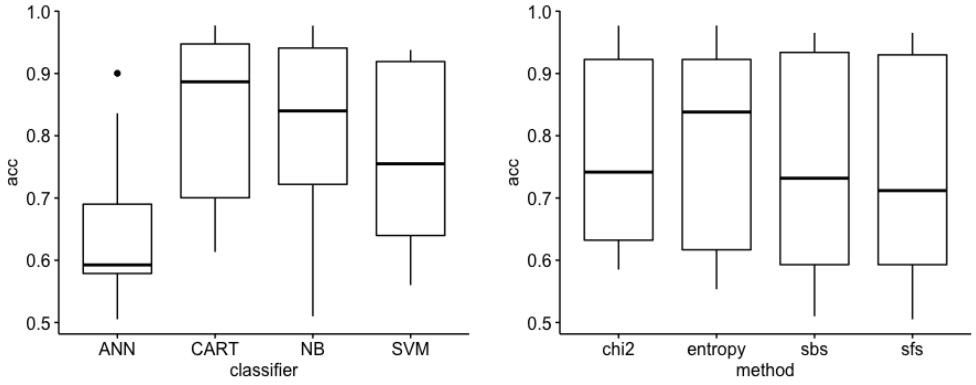
Table 4.1: ANOVA values of accuracy in relation to classifier, method and the interaction of classifiers and methods. RSS: Residual sum of squares. df: Degrees of Freedom. F: Mean Square for the Model divided by the Mean Square for Error (error/residual). $P(>F)$: the significance probability associated with F. The stars indicate the range of significant level: 0 "***" 0.001 "**" 0.01 "*" 0.05 " . " 0.1 " " 1.

The **-significance of Classifier concludes that the selection of classifier effect what accuracy is achieved. The variance of the each respective classifier is visualized in figure 4.1a. The box-plot entail that ANN on average perform worse than other classifiers. NB has the largest variance, with accuracy ranging from roughly 50% to 100%. CART performs the best, with the highest average accuracy.

The ANOVA result of Method conclude the selection of FS-method do not result in a significant difference in achieved accuracy. In figure 4.1b all methods manifests a large variance but roughly in the same range, which emphasizes the result of the ANOVA. That is, there is no significant difference in accuracy when applying different FS-methods.

Analyzing these results together, first they suggest that the selection of classifier impacts the expected accuracy. Secondly, which FS-method is used together with the classifier do not effect expected accuracy. Lastly, no unique combination of classifier and FS-method are statistically superior to others.

It's important to note these measurements only include classification with FS-selection, they do not enable any conclusions regarding the impact of using, versus not using feature selection. Therefore, that is the next point of interest.



(a) (b)

Figure 4.1: Variation in accuracy in respect to (a) classifiers and (b) FS-methods. Vertical axis represents accuracy. The horizontal bar in each box represents the mean accuracy. The box represents the range where 50% of the measurements was found. The vertical lines connected to the boxes includes represents the remaining 25% on each side. Dots represents outliers.

4.2 Classification Improvements

After performing classification with and without FS-methods, all results of each classifier was collected. Results are presented in tables for ANN 4.2a, CART 4.2b, NB 4.2c and SVM 4.2d where the highest achieved accuracy is highlighted in bold format. The improvement is measured in gain, the ratio between best achieved FS accuracy and full dataset accuracy.

ANN	MIAS	EN	RHH	WBCD
Chi2	0.58	0.59	0.64	0.71
Entropy	0.56	**0.84**	**0.90**	0.62
SBS	0.54	0.55	0.59	**0.74**
SFS	**0.59**	0.51	0.59	0.68
Full	0.57	0.68	0.60	0.53
Gain	0.04	0.24	0.51	0.41

1.2

(a)

CART	MIAS	EN	RHH	WBCD
Chi2	0.62	0.74	0.94	**0.98**
Entropy	0.61	0.84	0.94	**0.98**
SBS	0.70	0.70	0.93	0.96
SFS	0.70	0.70	0.94	0.97
Full	**0.75**	**0.95**	**0.95**	0.95
Gain	-0.06	-0.11	-0.01	0.03

-0.15

(b)

NB	MIAS	EN	RHH	WBCD
Chi2	**0.75**	0.75	0.93	0.97
Entropy	0.72	0.55	0.93	**0.98**
SBS	0.51	0.72	0.93	0.97
SFS	0.51	0.72	0.93	0.97
Full	0.57	**0.93**	**0.96**	0.96
Gain	0.30	-0.19	-0.03	0.02

0.1

(c)

SVM	MIAS	EN	RHH	WBCD
Chi2	0.59	0.75	0.92	0.66
Entropy	0.59	0.84	0.92	0.66
SBS	0.59	0.75	**0.94**	**0.94**
SFS	0.56	0.75	0.90	**0.94**
Full	**0.74**	**0.90**	0.64	0.64
Gain	-0.20	-0.07	0.46	0.45

0.64

(d)

Table 4.2: Mean accuracy of 10-fold cross-validation accuracies achieved on each dataset when applying some FS-method or with Full dataset. Tables are categorized by classifier; (a) ANN, (b) CART Decision tree, (c) Naïve Bayes and (d) Support Vector Machine. The highest accuracy is highlighted by bold font and Gain represents ratio between the best accuracy achieved by FS and the accuracy by Full dataset. Accumulated Gain is last on the Gain row. Best accuracy by FS is equivalent or greater than with Full dataset in every instance except in (b) with CART classifier on MIAS dataset.

4.2.1 Classification Improvements' Significance

To study the effect of feature selection on classification accuracy we perform a t-test. The t-test values constitutes of the accuracies achieved without feature selection as one distribution. The other are the best accuracies achieved with feature selection. The t-test values are summarized in 4.3.

Comparing the t-test results of all classifiers, t-stat value suggests a 28% increased performance when applying feature selection, see table 4.4. However, significance is insufficient to reject the null hypothesis that distributions are equal. Consequently we can not conclude feature

FS	MIAS	EN	RHH	WBCD
ANN	0.63	0.83	0.90	0.73
CART	0.67	0.83	0.92	0.97
NB	0.77	0.74	0.94	0.97
SFS	0.57	0.83	0.91	0.93

(a)

Full	MIAS	EN	RHH	WBCD
ANN	0.60	0.53	0.57	0.64
CART	0.77	0.69	0.90	0.94
NB	0.77	0.66	0.94	0.96
SFS	0.57	0.73	0.90	0.61

(b)

Table 4.3: (a) The highest accuracy achieved by any FS-method on all classifier-dataset combinations. (b) corresponding accuracies achieved without any feature selection on Full dataset. T-test comparing the distributions results in significantly higher results by classification with feature selection.

selection significantly improves classification accuracy based on our data.

To further analyze the differences, t-test is performed to compare each accuracy by each individual classifier, with and without feature selection. The tests concluded a significant improvement of accuracy using FS on ANN. No significant increase nor decrease in accuracy could be proven for the other classifiers; DT, CART and SVM.

Classifier	**t-stat**	**P**	
All	1.28	0.21	
ANN	2.55	0.04	*
CART	-0.43	0.68	
NB	0.04	0.97	
SVM	0.95	0.38	

Table 4.4: Significant values of t-test on Classifiers. t-stat is a ratio between the difference between two groups and the difference within the groups. P: Significance probability. The stars indicate the range of significant level: 0 "***" 0.001 "**" 0.01 "*" 0.05 " . " 0.1 " " 1.

The absence in the significance among the t-test results may be a consequence of data shortage. However, t-stat indicates some differences among classifiers that we'll analyze further.

4.2.2 Differences among Classifiers

We found in 4.1.1 that the selection of classifier affect the expected accuracy. In 4.2.1 t-test scores concluded benefit of feature selection varies between different classifiers. Ranking the accumulated gain of each classifier from tables 4.2, we construct table 4.5 which clarifies variation of improved accuracy between classifiers. We'll now look at each classifier in turn:

Rank	Classifier	Accumulated gain	Avg gain
1	Artificial Neural Network	1.2	30%
2	Support Vector Machine	0.64	16%
3	Naïve Bayes	0.1	0.3%
4	Decision Tree	-0.15	-0.4%

Table 4.5: Ranking of which classifiers gained most accuracy when comparing feature selection to full dataset.

Artificial Neural Network

Looking at the table 4.5 the accumulated gain was 1.2 which was the highest among all classifiers. However, ANN consistently performs the worst of all classifiers in terms of accuracy. The ANN also provides the least consistent results with strong fluctuations in the results and wide standard deviation margins. Such fluctuations may suggest issues regarding convergence as an effect of using default parameters in the network. These characteristics are evident in plot 4.2d showcasing the accuracy by each FS-method as a function of number of attributes. However, t-test show an significant increase in classification accuracy when feature selection is applied.

Support Vector Machine

SVM improves accuracy in two out of four datasets in table 4.2d. In these datasets the best performance is achieved by using wrapper methods.

The SVM behaves differently in respect to each dataset which may conclude different kernels of the SVM is needed on different datasets.

Improvements are seen with a larger subset of attributes in the EN dataset 4.5b. A negative trend on accuracy is observed on the WBCD dataset 4.5d which might indicate an issue with dimensionality. In 4.5c an improved accuracy is evident with maximal accuracy achieved on a subset suggesting a positive effect of feature selection.

Decision Tree

The decision tree shows consistent performance, generally increasing accuracy with an increased amount of features. Although in cases like plot 4.3a and 4.3b best accuracy is achieved with a subset of features displaying evident benefits of feature selection.

Naïve Bayes

In plot 4.4d the accuracy presents little to no improvement when increasing the number of attributes. In three out of four datasets NB produces the highest accuracy with very few attributes indicating benefits of feature selection.

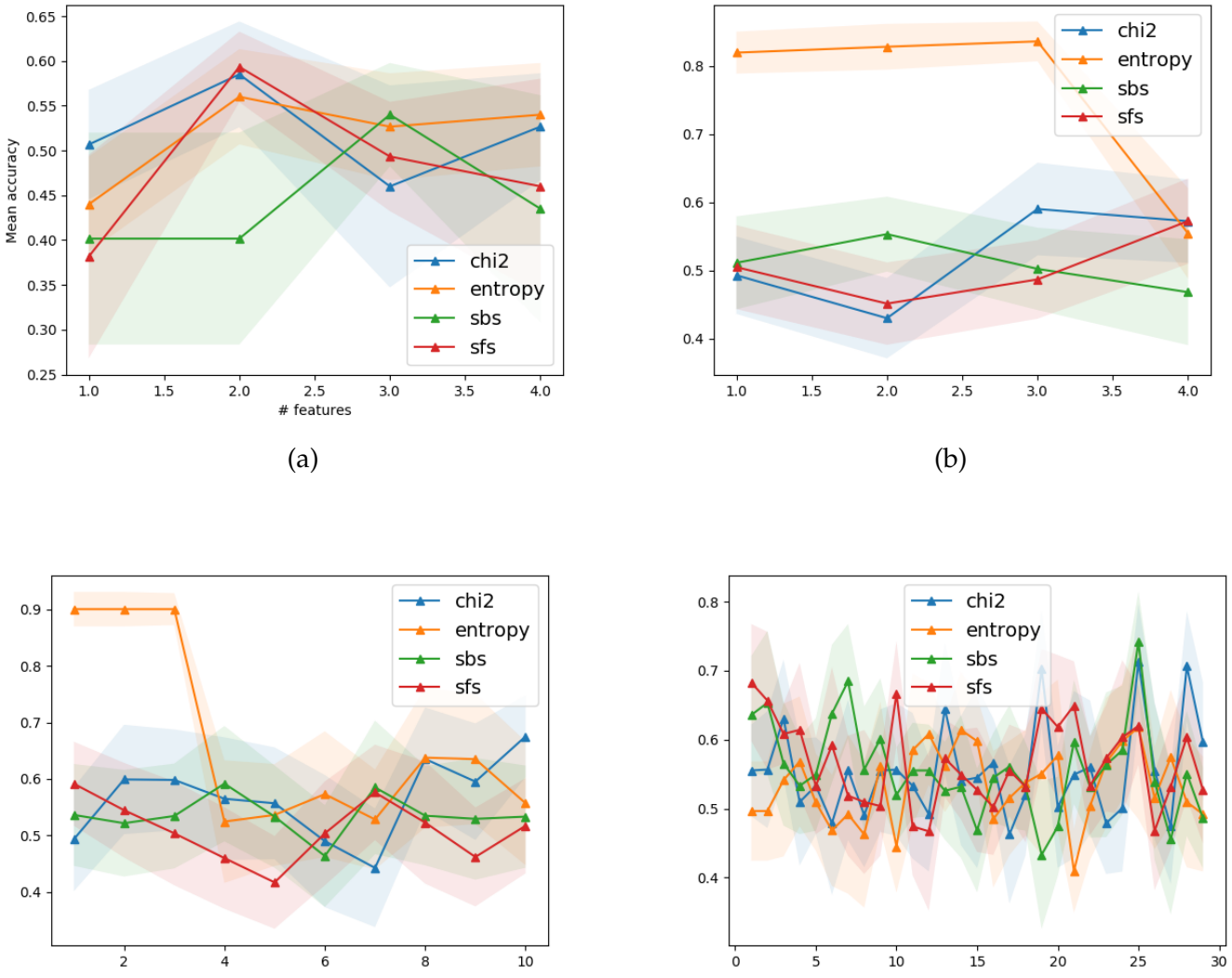


Figure 4.2: Classifier ANN. Each plot corresponds to datasets (a) MIAS, (b) EN, (c) RHH and (d) WBCD. x-axis is number of features, y-axis is mean accuracy achieved by corresponding feature selection method. Shaded area represent the standard error for each FS-method.

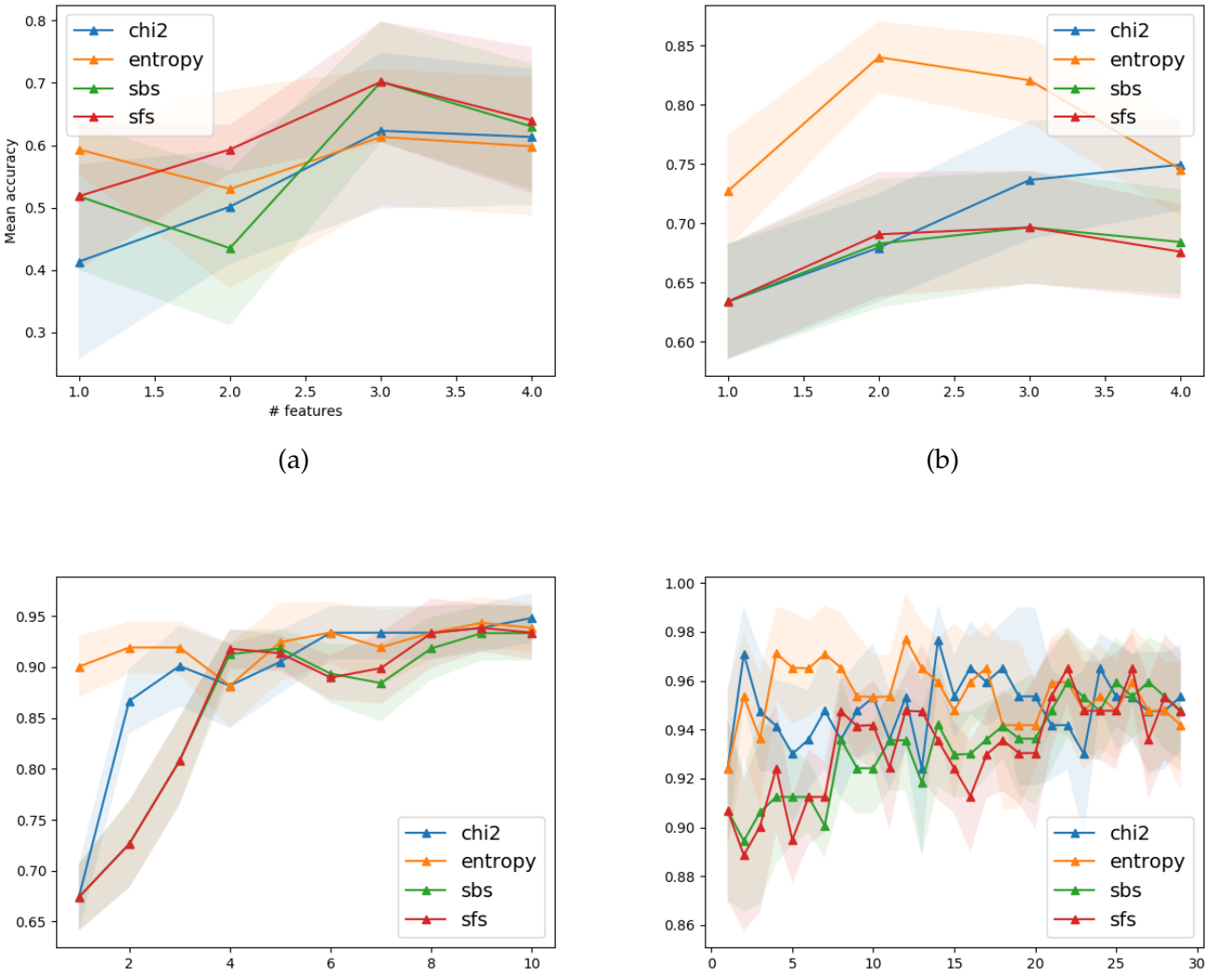


Figure 4.3: Classifier CART. Each plot corresponds to datasets (a) MIAS, (b) EN, (c) RHH and (d) WBCD. x-axis is number of features, y-axis is mean accuracy achieved by corresponding feature selection method. Shaded area represent the standard error for each FS-method.

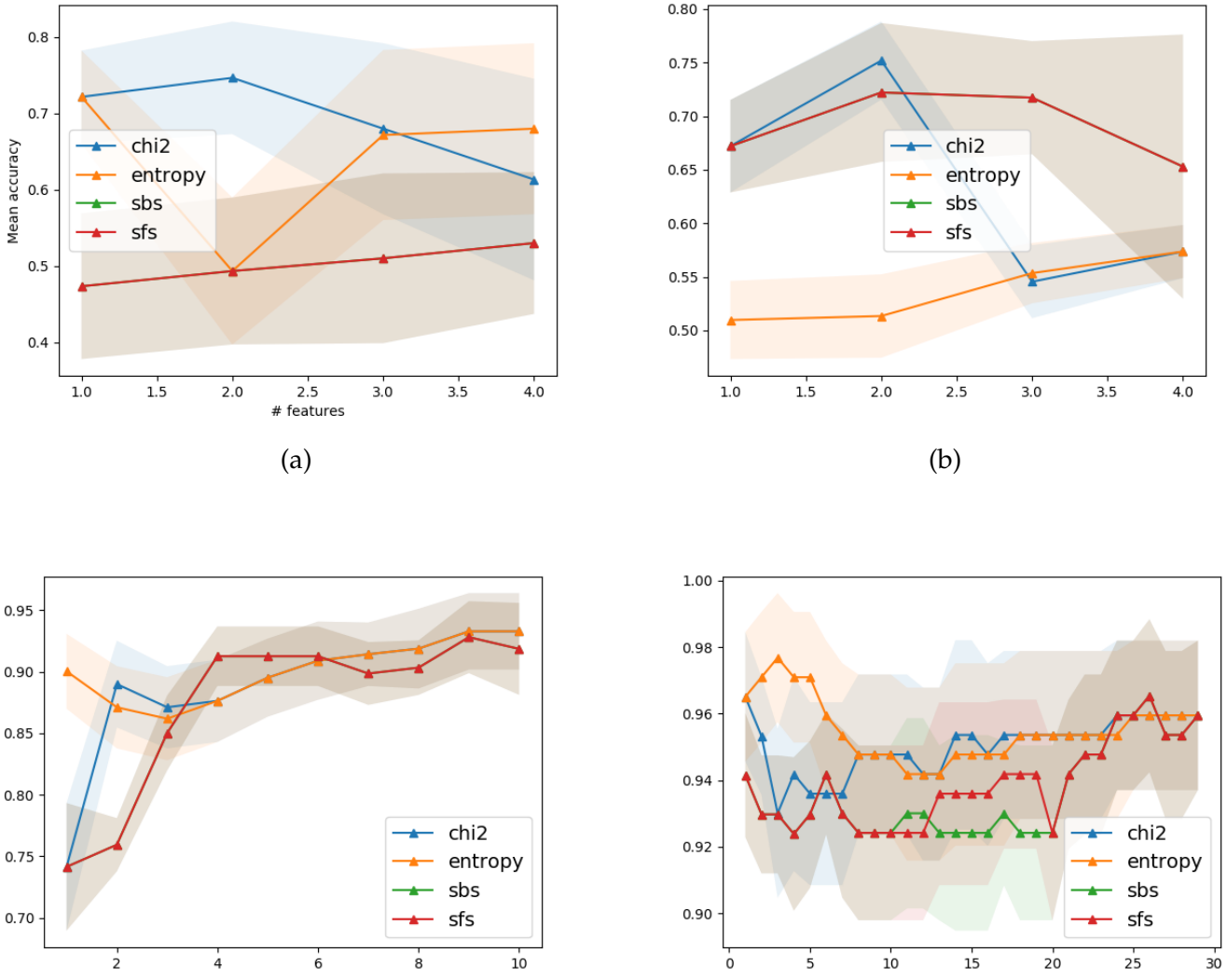


Figure 4.4: Classifier NB. Each plot corresponds to datasets (a) MIAS, (b) EN, (c) RHH and (d) WBCD. x-axis is number of features, y-axis is mean accuracy achieved by corresponding feature selection method. Shaded area represent the standard error for each FS-method.

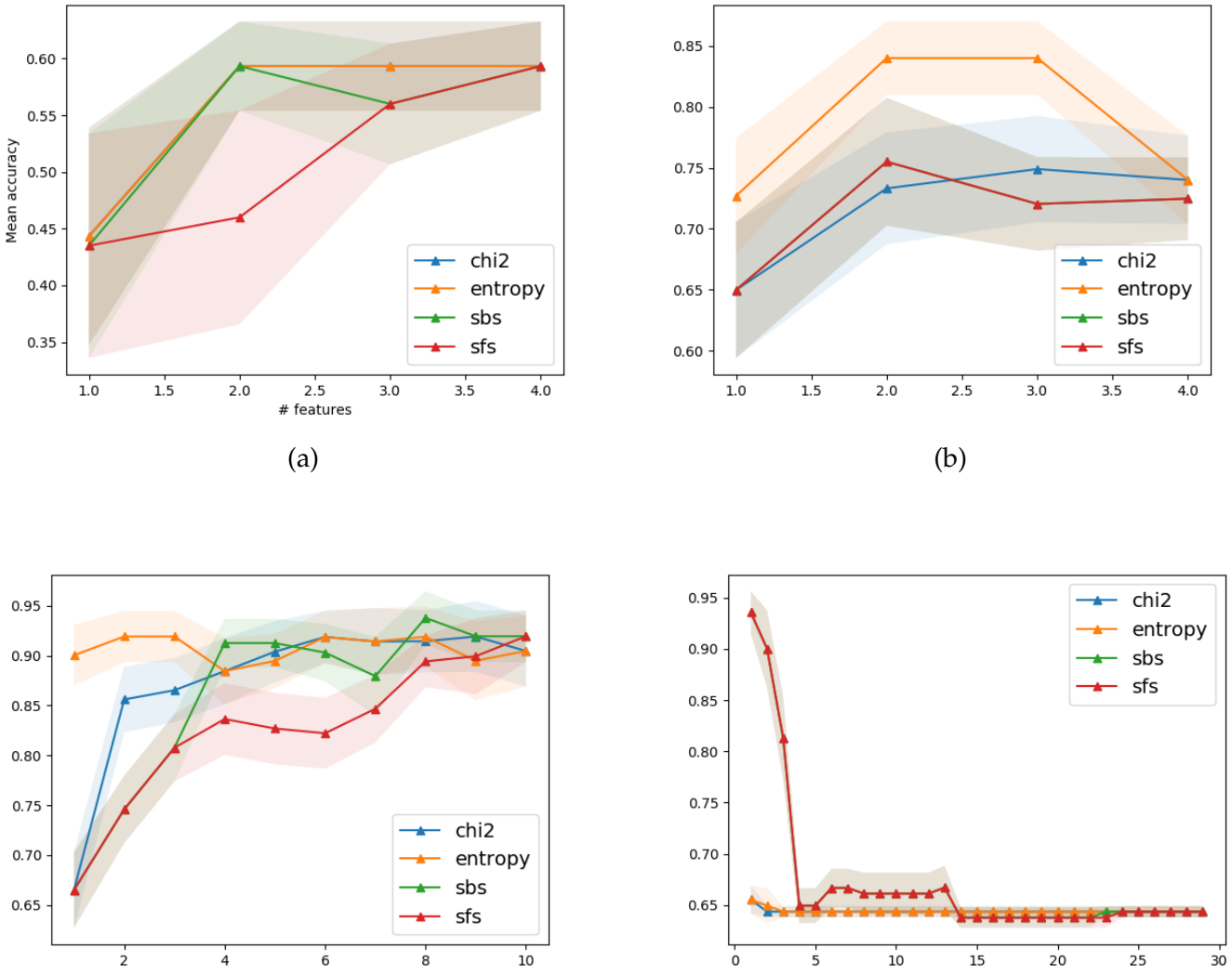


Figure 4.5: Classifier SVM. Each plot corresponds to datasets (a) MIAS, (b) EN, (c) RHH and (d) WBCD. x-axis is number of features, y-axis is mean accuracy achieved by corresponding feature selection method. Shaded area represent the standard error for each FS-method.

4.3 Computation time

Profiling the execution of running all experiments approximately 100% of CPU-time was allocated to the Wrapper algorithms as showed in table 4.6. Finding the best possible subset of features is considered a NP-hard problem meaning a solution can not be found in polynomial time. This clearly suggest favouring filtering methods when choosing a feature selection method having limited computational resources.

Function	**Cumtime**	**Ratio**
Chi2	3.759	0 %
Entropy	42.8	0 %
SBS	3751.879	13 %
SFS	24785.653	87 %

Table 4.6: Table of CPU usage for the four FS methods. Cumtime is the cumulative time spent in this and all subfunctions (from invocation till exit). Ratio is the percentage of the total run time.

Chapter 5

Discussion

In the experiments ANN, DT, NB and SVM classifiers was used to classify begin or malignant breast tumours. The classification accuracy was compared between using, or not using feature selection (FS). The FS-methods included in the tests were filter methods Chi2 and Entropy, and wrapper methods by SBS and SFS.

The results could not prove there was a significant increase in classification accuracy using FS, when considering all classifiers together. However, applying FS combined with ANN, we found a significant increase in accuracy compared to using ANN without FS. It leaves the conclusion that the effect of feature selection is dependent on which classifier is used.

5.1 Influence of Feature Selection

As stated in the results, all classifiers but CART had an indication of improved classification accuracy when applying feature selection. However, only ANN had a statistically significant increase. In our experiments ANNs increased its accuracy on all datasets by using FS. The benefits applying FS to ANNs consists of potentially lessened effort gathering, and processing attributes at data collection. It can also be considered a reguralization method which may prevent overfitting. Another benefit is lessened computation time, as found in the results section 4.3. As stated by Martei et al. [19], there is a large demand of a

more streamlined and efficient process when it comes to breast cancer classification. Feature selection seems to offer such benefits to ANNs in this process.

Analytical tests of the results showed choosing the suitable classifier for the dataset rendered the largest effect on accuracy. When it comes to the tho groups of feature selection methods, filter and wrappers, no significant difference presented itself. Still, the choice of one or the other poses a dilemma. As filter methods proved computationally fast in comparison to wrappers, many various filter methods can be evaluated efficiently to find a good subset. But, wrappers sometimes provided better result, although they manifested a heavily prolonged computation time at training as showed in table 4.6.

If wrapper methods are to be used there may be a large benefit in constructing a search approach for the NP-hard part of the problem, instead of evaluating the full search space. Such a study has been conducted by Panthong and Srivihok [21] and improved both classification accuracy, and reduced runtime. Another approach is not searching the through all possible subset, with some domain knowledge a more narrow span of parameters could be manually picked which in turn restricts the number of computations.

Using default classifiers might cause the observed fluctuation and variance in the results. A classifier with n attributes might need very different parameters than the same classifier with $n + 1$ parameters to achieve optimal performance. This raises an important question, should a a classifier be tuned and optimized before applying feature selection, after or during features are selected. Before may raise the problem mentioned above, during logically offer the best results but in some cases infeasible computation time and after may miss feature subsets which could have performed better if search had been made with different parameters.

5.2 Comparing Classification Accuracy

Reports that achieve high classification accuracy such as Akay [2] decide on one classifier and FS-method and optimize the parameters of the classifier to both the FS-method and the dataset. It results in high

performance but leaves the question how such classifier and FS combinations should be chosen and how they perform on other data.

On average FS improved the result of the ANN classifier by 30%. This result can be compared with the 28% gain achieved by Karabulut, Özel, and İbrikçi [17]. The similar result can be explained by use of similar classifiers and although Karabulut, Özel, and İbrikçi [17] used different filter methods than this study in line with our findings the FS-method is not a deciding factor for accuracy. The SVM classifier received an average gain of 16%. Comparing this results with Babaoglu, Findik, and Ülker [5] our result shows a greater impact of FS using a SVM classifier. This is probably an effect of different datasets and FS methods being used as both these factor effect the accuracy. However, our result for improved classification accuracy using SVM is not significantly proven as mentioned in section 4.2.1.

5.3 Further Research

Further investigation of a methodology of finding the best possible classifier and FS-method.

As mentioned in the chapter 5.2 an important question to be answered is wether a high performing classifier and FS-method be found first then optimized by tuning or it it the other way around.

More recent strategies of diagnosing breast cancer involves sampling microRNA from patients. Other diseases have been diagnosed with by this strategy and presented promising results. As a sample of microRNA contains around 2 000 features, selection may offer a huge benefit in line with our findings that a increased number of feature benefits more from feature selection.

5.4 Effect of Limitations

Due to the limited amount of breast cancer datasets found and utilized in the study, it is difficult to confidently draw conclusions regarding all breast cancer classification at large as dataset may vary from the ones used in this report.

Limited resources has also resulted in a reduced subset of FS-methods studied. While having two methods of each FS-family, filters and wrappers, there are many more which may have produced different results than we have achieved.

5.5 Ethical Aspects

The best classification accuracy found in this report was 97%. Studies show machine learning already outperforms medical experts in setting a correct diagnose of breast cancer [14]. As diagnostics develops from being made by humans to machines, many factors need to be considered. Do humans trust computers enough to allow this development, should they be informed their diagnosis is set by algorithms? If so, how can we explain a certain output when many algorithms are truly hard to interpret. Lastly, what data should be used for training, only collecting data of those who have access to such healthcare may introduce a bias against other demographics of the population.

The data used in this report origin from real patients that may be experience discomfort during mammographies, FNA sampling or with other method was used when collecting the data. The data also holds sensitive information ruling the patients future health. Data is to our knowledge never collected without a patients consent and carries no information that can allows any identification of the patient.

5.6 Sustainability

We trust the reliability in our findings and believe they contribute to the accumulated knowledge of the field as they are made available. In that sense the of classification and breast cancer research progresses forward and can in turn make new discoveries that enables a more sustainable future.

5.7 Retrospective

While we perceive the basis and conduction of our approach investigating our research to be solid, would we do it again slight changes would be made. The report has a wide scope covering both breast cancer and feature selection. The latter is affected by many variables such as dataset, classifier and FS-method. Shifting focus to one of the areas would allow for deeper analysis. In the case of breast cancer more domain knowledge could be studied such as what attributes actually are important. Looking at only feature selection the supply of datasets would be larger when not restricted to breast cancer and thus offer more material for comparisons of classifier-FS-data interaction.

Chapter 6

Conclusion

Applying feature selection methods provides an improved classification accuracy on benign or malignant breast cancer when using an Artificial Neural Network classifier. The improvement of ANN was consistent over multiple datasets. No correlation between accuracy achieved by ANN and what FS-method used was found suggesting it is case-to-case dependent.

When using classifiers Decision Tree, Naïve Bayes and Support Vector Machine no increase, or decrease by using feature selection is significantly evident over multiple datasets. In some observations these classifiers manifested increased classification accuracy with feature selection compared to using the full dataset. The feature selection methods that improve these methods the most are generally wrappers such as SFS and SBS although they demand large computational time compared to filter methods, see results 4.3.

The machine learning classifier that overall benefited most from feature selection in terms of accuracy was Artificial Neural Network. On four different datasets with significantly different compositions, accuracy of ANN with feature selection compared to no feature selection, was improved. The largest improvement was a 51% increase when applying the feature selection method entropy.

Bibliography

[1] Layla Abdel-Ilah and Hana Šahinbegović. "Using machine learning tool in classification of breast cancer". In: *CMBEBIH 2017*. Ed. by Almir Badnjevic. Singapore: Springer Singapore, 2017, pp. 3–8.

[2] Mehmet Fatih Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". In: *Expert Systems with Applications* 36.2, Part 2 (2009), pp. 3240–3247.

[3] Michelle D Althuis et al. "Global trends in breast cancer incidence and mortality 1973–1997". In: *International Journal of Epidemiology* 34.2 (2005), pp. 405–412.

[4] Edoardo Amaldi and Viggo Kann. "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems". In: *Theoretical Computer Science* 209.1 (1998), pp. 237–260.

[5] Ismail Babaoglu, Oğuz Findik, and Erkan Ülker. "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine". In: *Expert Systems with Applications* 37.4 (2010), pp. 3177–3183. ISSN: 0957-4174.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[7] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers and Electrical Engineering* 40.1 (2014), pp. 16–28. ISSN: 0045-7906.

[8] Jie Zhi Cheng et al. "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans". In: *Scientific Reports* 6 (Apr. 2016), pp. 2045–2322.

[9] A.M. Culpan. "Radiographer involvement in mammography image interpretation: A survey of United Kingdom practice". In: *Radiography* 22.4 (2016), pp. 306–312.

[10] Walter Daelemans et al. "Combined optimization of feature selection and algorithm parameter interaction in machine learning of language". In: *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*. Berlin: Springer, 2003, pp. 84–95.

[11] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[12] James A. Adams George Dimitoglou and Carol M. Jim. "Comparison of the C4.5 and a Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability". In: *Journal of Computing* 4 (8 2012).

[13] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *An Introduction to Variable and Feature Selection* 3 (2003), pp. 1157–1182.

[14] Dr Robert F Harrison and Dr Simon S Cross. *Fine Needle Aspirate of Breast Lesions Dataset*. 1993. URL: `http://www.phil.gu.se/ann/data/`.

[15] Kononenko Igor. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23 (1 2001), pp. 89–109.

[16] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 1461471370, 9781461471370.

[17] Esra Mahsereci Karabulut, Selma Ayşe Özel, and Turgay İbrikçi. "A comparative study on the effect of feature selection on classification accuracy". In: *Procedia Technology* 1 (2012), pp. 323–327.

[18] M. Li and Z. H. Zhou. "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37 (Nov. 2007), pp. 1088–1098.

[19] Yehoda M. Martei et al. “Breast Cancer in Low- and Middle-Income Countries: Why We Need Pathology Capability to Solve This Challenge”. In: *Clinics in Laboratory Medicine* 38.1 (2018), pp. 161–173.

[20] Akin Ozcift and Arif Gulten. “Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms”. In: *Computer Methods and Programs in Biomedicine* 104.3 (2011), pp. 443–451.

[21] Rattanawadee Panthong and Anongnart Srivihok. “Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm”. In: *Procedia Computer Science* 72 (2015). The Third Information Systems International Conference 2015, pp. 162–169. ISSN: 1877-0509.

[22] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[23] Raúl Ramos-Pollán et al. “Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis”. In: *Journal of Medical Systems* 36 (Aug. 2012), pp. 2259–2269.

[24] Sebastian Raschka. “MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack”. In: *The Journal of Open Source Software* 3.24 (Apr. 2018).

[25] Lars Stahle and Svante Wold. “Analysis of variance (ANOVA)”. In: *Chemometrics and Intelligent Laboratory Systems* 6.4 (1989), pp. 259–272. ISSN: 0169-7439.

[26] László Tabár et al. “Beyond randomized controlled trials - Organized mammographic screening substantially reduces breast carcinoma mortality”. In: 91 (June 2001), pp. 1724–31.

[27] Manolis Wallace. *Emerging Artificial Intelligence Applications in Computer Engineering*. Amsterdam, BG: IOS Press, 2007.

[28] Maoxin Wu and David E. Burstein. “Fine Needle Aspiration”. In: *Cancer Investigation* 22.4 (2004), pp. 620–628.

Appendix A

Appended Material

A.1 Classifier parameters

Classifier parameters		
ANN: Multi-layer Perceptron	Hidden layers	2
	Layer size	100
	Activation	ReLU
	Solver	Adam
	alpha	0.0001
	batch size	auto
	learning rate	constant
	learning rate init	0.001
	power t	0.5
	tolerance	1e-4
	beta1	0.9
	beta2	0.999
	epsilon	1e-8
Decision Tree: CART	Criterion	Gini
	Splitter	Best
	Max depth	None
	min samples split	2
	min samples leaf	1
	min weight fraction leaf	0
	max features	None
	random state	None
	max leaf nodes	None
	min impurity decrease	0
	class weigh	None
	presort	False
Naive Bayes	Type	Gaussian
	Priors	None
SVM	Kernel	Rbf
	Gamma	auto
	Penalty	1.0
	Shrinking	True
	Tolerance	1e-3

Table A.1: Most defining parameters of each classifier used