Evaluating feature selection in a marketing classification problem

Degree project
Abstract

Nowadays machine learning is becoming more popular in prediction and classification tasks for many fields. In banks, telemarketing area is using this approach by gathering information from phone calls made to clients over the past campaigns. The true fact is that sometimes phone calls are annoying and time consuming for both parts, the marketing department and the client. This is why this project is intended to prove that feature selection could improve machine learning models.

A Portuguese bank gathered data regarding phone calls and clients statistics information like their actual jobs, salaries and employment status to determine the probabilities if a person would buy the offered product and/or service. C4.5 decision tree (J48) and multilayer perceptron (MLP) are the machine learning models to be used for the experiments. For feature selection correlation-based feature selection (Cfs), Chi-squared attribute selection and RELIEF attribute selection algorithms will be used. WEKA framework will provide the tools to test and implement the experiments carried out in this research.

The results were very close over the two data mining models with a slight improvement by C4.5 over the correct classifications and MLP on ROC curve rate. With these results it was confirmed that feature selection improves classification and/or prediction results.

Keywords: Neural networks, term deposit, bank marketing.
Preface

I want to thank my family who has been important support for me all along my life and especially in this important step in my professional career, my father who has done all this possible and provided me with the resources not only economically but moral and intellectually. My mother who made me a person of values and showed me honesty and knowledge empowers you. My brother and sister who have set the path for me to walk. Also thank all the persons whom provided the information, time and support for this project, my supervisor who always showed available and the people who guided me with advice.

It has been a long and challenging way till now that I could not have been able to make without support and values taught to me since I was younger.
Contents

1 Introduction ___________________________________________________________ 5
   1.1 Problem Introduction ___________________________________________ 5
   1.2 Previous research _____________________________________________ 5
   1.3 Problem definition ______________________________________________ 6
   1.4 Purpose and research question/hypothesis __________________________ 6
   1.5 Scope/limitation ________________________________________________ 6
   1.6 Target group ___________________________________________________ 6

2 Background ___________________________________________________________ 7
   2.1 Theory Background ______________________________________________ 7
   2.2 Neural Networks in Marketing ______________________________________ 9

3 Method ________________________________________________________________ 10
   3.1 Scientific approach ______________________________________________ 10
   3.2 Data _____________________________________________________________ 10
   3.3 Framework ______________________________________________________ 14
   3.4 Analysis _________________________________________________________ 15
   3.5 Reliability _______________________________________________________ 19
   3.6 Ethical considerations ___________________________________________ 20

4 Results/Empirical data ____________________________________________________ 20
   4.1 Approaches for reporting results ____________________________________ 20
   4.2 Results Analysis ________________________________________________ 21

5 Results Conclusion______________________________________________________ 24

6 Discussion ______________________________________________________________ 25
   6.1 Problem solving/result ____________________________________________ 25
   6.2 Method reflection ________________________________________________ 25

7 Conclusion ______________________________________________________________ 26
   7.1 Conclusions ______________________________________________________ 26
   7.2 Further Research ________________________________________________ 27
1 Introduction

In this section a short description of the problem, of previous work and the scope of this research is presented.

1.1 Problem Introduction

The goal of this research and experiments is to compare the effects of different feature selection methods over data mining algorithms. This is done using a data set from a Portuguese bank that gathered telemarketing registries through phone calls for obtaining term deposit contracts. This is carried out with two machine learning algorithms to demonstrate if feature selection is an important step and its effects in data mining. Feature selection is a process where an algorithm determines the utility or importance of an attribute on the data set. The classifier or machine learning algorithm can then exclude attributes that are not important, or assign different weights according to how important different attributes are.

1.2 Previous research

Machine learning or more specific artificial neural networks has been broadly used in sales prediction. One example is to improve sales and reduce return quotes using Bayesian learning and top down search algorithms for newspaper retailers [1]. In another project Support Vector Machines (SVM) was used for predicting the amount of magazines that should be delivered to newly opened stores [2].

The probably most common example is web sites that classifies visits and search terms to offer recommendations according to that historical data.

The data used in this research has been used in previous research where the goal was to show a data mining approach for predicting the success of marketing phone calls through four models: logistic regression, decision trees, neural networks and support vector machines. In that research area under the curve (AUC) and area of the LIFT cumulative curve were used as metrics to provide credibility and precision of the tests [7].

Moreover the experiments were carried together with manual and automatic feature selection leaving the data set with 22 features from the
original 150 features. Manual selection consisted of problem domain knowledge to define a set of questions to select the most useful features.

1.3 Problem definition

The goal of this thesis is to compare different feature selection algorithms and how effective they are on a number of different machine learning algorithms.

1.4 Purpose and research question/hypothesis

RQ1. Can feature selection improve accuracy and performance of machine learning algorithms on the selected dataset?
RQ2. How accurate is machine learning algorithms for predicting if a person will become a customer of the bank using the selected dataset?
RQ3. Which machine learning algorithm benefits the most of feature selection?

In order to answer these questions machine learning models together with feature selection will be used and tested. These questions are the basis of this thesis and it is the purpose to give an answer and explain them.

1.5 Scope/limitation

The goal of this thesis is not to implement a complete system, but rather investigate the possible improvements of feature selection for decision trees and neural networks. It will also find the best approach or combination of these two.

1.6 Target group

This report is intended for the scientific community that can use this knowledge as a base for developing a complete system including feature selection processes as an important and basic step in data mining. Also the public in general who is interested in data mining processes and statistical analysis.
2 Background

In this chapter the problem background and why this is relevant to the target group referred to above is presented.

2.1 Theory Background

Neural networks have been attracting computer scientists, engineers and mathematicians since they appeared with the concept of being an adaptive system or universal functional approximator [3].

It all started with the McCulloch & Pitts model where the neurons have a weighted input and the sum of all the weighted inputs and a preset threshold determines if the neuron is activated or not. Then the Mark I Perceptron came which was based on McCulloch & Pitts model but also adds some fixed pre-processing to work on basic pattern recognition [4]. The problem with single layer perceptrons was the inability to do some more complex pattern recognition operations. After this, several models came along like Adaptive Resonance Theory (ART), Self-Organizing Maps (SOM) and multilayer Perceptrons which were able to approximate more complex functions.

Applications goes from physics, finances, business sales, marketing, economics or even social science.

Neural networks are basically an imitation of the human brain where the cells are massively interconnected in between to form a parallel network of communication. This is why neural networks are useful for complex computation or exhaustive information processing in contrast to conventional machines that works sequentially.

Moreover neural networks are similar to the human brain in that they learn by example. This is they have to be trained in order to be able to learn from patterns or any kind of information. There exist different neural networks models and an example of the basic layout is shown in Figure 2.1. All the layouts share some basic characteristics:

- Learning: The network adapts depending on the information to offer the desired output.
- Generalization: Once the network is trained it can handle slight changes in the information or the way it works.
- Massive Parallelism: The neurons in the network are interconnected in between to communicate with each other.
• Fault-tolerance: Even if some of the elements suffer alterations it can keep working since all the elements in it are parallelized.

Figure 2.1 Multi-layer neural network model.

One way of classifying neural networks is in the following two types:
• Supervised Learning: Includes a teacher in the learning process which means there is something that supervises the learning process for the output so that the network measures the error according to the desired output (teacher) and the system’s output, then adjusts to offer closer or more accurate results. In this experiments this is the used approach for the machine learning algorithms. Figure 2.2 shows the structure of this type.

Figure 2.2 Overview of supervised learning [3]
• Unsupervised Learning: The learning process is carried away in a self-organizing way meaning that it does not have a teacher as supervised learning, also there is no external entity to adjust the weights and the desired output is not available during the training phase. On the other way there are competitive neurons that compete for the best representation of the input pattern.

![Figure 2.3 Overview of Unsupervised Learning [3]](image)

2.2 Neural Networks in Marketing

Sales are the basis of every institution either it sells a product or a service, sales provide the income for the company’s sustainability. This is why neural networks have become an interesting topic in this area because this kind of technology has the ability to predict, cluster or identify patterns in any kind of information which, nowadays, either it is text, numeric or any type of information can be processed with a number of algorithms and/or models to provide forecasting and also assure that the results are accurate.

Forecasting consists of learning one or more variables from other already known variables derived from previous knowledge [6]. However forecasting is not only about selecting data from the past and expect accurate prediction results. It mostly depends on the selected data from the past to fit new data and make predictions, classifications, etc. for the future.

Learning models work differently and can result in distinct accuracy values for the same data set, which is why the goal of this report is to compare different algorithms and select the most correct algorithm to offer an alternative in marketing strategies.
Since sales are volatile, affected by external issues or variables not contemplated in the past, prediction needs to use a non-linear model to give a better approach and accuracy over the unstable data [6].

Having this context can be noticed that neural networks are innovating and new applications are flourishing, it is just a matter of creativity to implement them either in physics, business or any other applications [3].

3 Method

In this chapter the used dataset, motivation for preprocessing and selecting the neural network algorithm and an interpretation of the obtained results are described.

3.1 Scientific approach

The goal of this thesis is the comparison of the effects and possible improvements of feature selection in machine learning algorithms. The comparison is carried out through experiments and measurements of performance and results, providing quantitative data that will help in the understanding of the mentioned processes and their functionalities.

3.2 Data

The data used in this research comes from the UCI machine learning repository and was gathered by the Portugal Bank from telemarketing campaigns and contains 21 attributes including the class value, which is the value to predict (Subscripted or not) [7]. Table 3.1 below shows a description of each of the attributes gathered in the database:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Persons age</td>
<td>Numeric</td>
</tr>
<tr>
<td>Job</td>
<td>Type of job</td>
<td>Nominal</td>
</tr>
<tr>
<td>Marital</td>
<td>Marital status</td>
<td>Nominal</td>
</tr>
<tr>
<td>Education</td>
<td>Education level</td>
<td>Nominal</td>
</tr>
<tr>
<td>Default</td>
<td>Has credit in default</td>
<td>Nominal</td>
</tr>
<tr>
<td>Housing</td>
<td>Has a housing loan</td>
<td>Nominal</td>
</tr>
<tr>
<td>Loan</td>
<td>Has a personal loan</td>
<td>Nominal</td>
</tr>
<tr>
<td>Contact</td>
<td>Communication type</td>
<td>Nominal</td>
</tr>
<tr>
<td>Month</td>
<td>Last contact month</td>
<td>Nominal</td>
</tr>
<tr>
<td>Day of week</td>
<td>Last contact day of</td>
<td>Nominal</td>
</tr>
<tr>
<td>Attribute</td>
<td>Description</td>
<td>Type</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>week</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>Duration of last contact</td>
<td>Numeric</td>
</tr>
<tr>
<td>Campaign</td>
<td># contacts in this campaign</td>
<td>Numeric</td>
</tr>
<tr>
<td>Pdays</td>
<td># days since last contact in past campaigns</td>
<td>Numeric</td>
</tr>
<tr>
<td>Previous</td>
<td># previous contacts in past campaigns</td>
<td>Numeric</td>
</tr>
<tr>
<td>Poutcome</td>
<td>Outcome of past campaigns</td>
<td>Nominal</td>
</tr>
<tr>
<td>Cons price idx</td>
<td>Consumer price index</td>
<td>Numeric</td>
</tr>
<tr>
<td>Euribor3m</td>
<td>Euribor 3 month rate</td>
<td>Numeric</td>
</tr>
<tr>
<td>Nr employed</td>
<td>Number of employees</td>
<td>Numeric</td>
</tr>
<tr>
<td>Subscribed(y)</td>
<td>Subscribed or not(target class)</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

Table 3.1 Information taken from [7]

3.2.1 Selection and pre-processing

Since the dataset is public for research it has already been used with different approaches or purposes and the attributes it comes with are diverse. This research differs from previous research in that it analyzes if all attributes are necessary or if it depends on the kind of application one wants to give to it. Using this data set is useful to measure the probabilities of a person to become a client and purchase a term deposit held at a financial institution (in this case the Portuguese bank) for a fixed term with retributions over the time.

Attributes selection is a highly important activity in machine learning. Large amounts of information is gathered and it is getting cheaper to accumulate and store it. It is important to store the important parts of the data that is correlated in certain ways and not just all the information available that maybe tomorrow will become useless or meaningless in a subject or task.

Selecting and filtering information is a time consuming task and even though different machine learning algorithms can offer similar results, feature selection can also have a large impact on the results [8].

The used dataset has already been filtered for privacy concerns and as can be seen from [7] only 22 out of 150 available attributes were chosen by semi-automatic methods. Some of the attributes were chosen manually for their
relevance in experience to support the decision and some other computationally chosen by their influence in the results as well. It is important to note this since the goal here is to compare different feature selection methods and their effects over different machine learning algorithms.

The framework used in this experiments is WEKA, explained later in this report Chapter 3.3. In WEKA there are several algorithms to determine dependency and influence of parameters in a dataset. All of them are basically based on three steps: the starting point to start searching, the next subset to evaluate and a stop criterion. A subset is composed of a number of attributes from the original dataset and the search algorithm evaluates each of these combinations [9].

In this research correlation-based feature selection (Cfs SubsetEval) is going to be tested. This algorithm determines dependency between attributes and the class, eliminating redundant and irrelevant data which can improve results and performance on a dataset [8].

The second attribute selection to be compared is RELIEF attribute evaluation which basically is based on nearest neighbor search. Finally the third algorithm is Chi squared attribute evaluation which evaluates the relevance of an attribute according to statistic values with respect to the class to determine its independence.

3.2.1.1 Correlation-based feature selection

This feature selection algorithm works with correlation between attributes in terms of the class (predicted attribute) and determines how dependent the attributes are between them and the class to eliminate redundant and irrelevant information, which is when two attributes cover the same predictive ability or they have no relation to the class.

Equation 3.1 shows the basis of this process:

\[ r_{cz} = \frac{r_{zi} - r_{ii}}{k} \]

Where \( r_{cz} \) is the correlation between the attributes and the predicted class, \( k \) is the dimension of the dataset, \( r_{zi} \) is the average of the correlation between attributes and the predicted class and \( r_{ii} \) is the average inter-correlation between attributes [8].

For this feature selection it is necessary to standardize all attributes to have a common basis to test co-relationship and treating attributes using discretization.
3.2.1.2 RELIEF attribute selection

RELIEF proposed by Kira and Rendell [10] is based on attribute estimation, which means assigning a weight or a relevance value to each attribute and the values that exceed a threshold are chosen. This algorithm is in essence a nearest neighbor search to a value \( V_1 \) of each class, evaluates the relevance of all the attributes and accumulates in \( W(f) \) with Equation 3.2. The nearest neighbor from the same class is a “hit” or a success otherwise it is a failure of class \( C \). At the end the final \( W(f) \) is divided to obtain the average between \([-1,1]\).

\[
W[f] = W[f] - \text{diff}(f, E_1, H) + \sum_{C \neq \text{class}(E_1)} P(C) \times \text{diff}(f, E_1, M(C))
\]

Equation 3.2

In general attributes gain relevance when a pair of values are the same and of different class. On the other hand attributes lose relevance when having different values and same class on pairs [10].

Another point to remark about the dataset is the unbalanced percentage per class instances. From a total of 41188 instances there are only 4640 of ‘yes’ class and the rest are ‘no’ class, which can cause the results to be uncertain in some cases or imprecise, this is why WEKA implements some methods of preprocessing for balancing classes. Other methods are presented in [10] but none are used in this research.

3.2.1.3 Chi-Squared attribute evaluation

This algorithm evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. The chi-square statistic measures the independence of two events providing a statistical test for asserting whether an association exists between two values. Equation 3.3
shows how chi-square statistic is calculated where $O_i$ is the observed frequency of the $i^{th}$ value and $E_i$ is the expected frequency of the $i^{th}$ value.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Equation 3.3 Chi-square equation.

This value determines that two events A and B are independent if $P(AB)=P(A)P(B)$ or $P(A|B)=P(A)$ and $P(B|A)=P(B)$ this means how much expected counts of ‘E’ and observed counts ‘O’ deviate from each other.

A large value of $\chi^2$ indicates that there is no association between the two values (observed counts and expected counts differ).

3.3 Framework

The framework used in this project is WEKA from the University of Waikato New Zealand. It is an open source framework written in Java that offers several algorithms for classification, regression, association and clustering data in different formats and sources. It also offers data pre-processing for different data types and purposes.

The framework can be used in two ways:

- CLI: recommended for in-depth analysis since it offers some functionality the GUI does not have and utilizes less memory for exhaustive processing.
- GUI: has three main options:
  - Explorer: Environment for exploring data.
  - Experimenter: Environment for statistical tests between learning schemes.
  - Knowledge Flow: Drag and drop interface for exploring and processing data.

WEKA organizes datasets in instances which are single data points consisting of one or more attributes.

Weka supports four main data types:

- Nominal: A defined list containing names as values, e.g. list of names of plants.
- Numeric: Real or integer values.
- Date: For treating date attributes with format like Java.
• String: Consecution of alphanumeric characters enclosed in “quotes”.

Machine learning algorithms in this framework are called classifiers independently of its purpose (classification, regression, clustering or association) and it also comes with filters for pre-processing data in runtime with many options and file formats for loading and saving datasets.

3.4 Analysis

In this research two basic algorithms among with three feature selection methods are used to compare the results and notify if there could be a noticeable difference with and without feature selection algorithms. The three basic algorithms to be used are as follows:

3.4.1 Decision Tree (J48 on WEKA or C4.5)

C4.5 is a model based on ID3 with some improvements for overfitting. The nodes of the tree contains the attributes from the input vector of the data set and the connections between them are values taken from the attribute value. The tree is built top-down, this way each leaf is a class prediction.

In order to decide which attributes to split at each node, the algorithm tests which attribute that best splits the data set into classes, this means nodes that do not contain a mix of both classes.

To measure which attribute best splits the class, the model calculates the entropy of a subset created for each available attribute. Entropy is the degree of randomness in the created subset and is calculated with Equation 3.4.

\[ H(s) = -\sum p(x) \log_2 p(x) \]

Equation 3.4 System entropy

\[ IG = H(S) - [(ss/\bar{d}s) \times H(c)] \]

Equation 3.5 Information Gain
H(S) is the complete system entropy value and H(c) is the child node entropy value multiplied by average value of the child node, this is ss the subset size divided by ds the data set size. Therefore the higher IG is the model learns more by splitting the selected attribute.

Basically this is how C4.5 works and everything is done by WEKA automatically however in some cases parameters could be changed to avoid overfitting, which means the tree tries to fit exactly the attributes values creating very specific rules for every case.

3.4.2 Multilayer perceptron

Multilayer perceptron is a feedforward neural network model composed of neurons called perceptrons. Perceptrons are units that receives, processes and outputs data depending on which layer they are located at. The units that receive information are located in the input layer and these units only receive the attributes from input data and send the information to the hidden layer where processing units are located, and then finally information is sent to the output layer. A model of a perceptron is shown in Figure 3.1.

As seen in figure above the input vector \{x_1, x_2, \ldots, x_n\} is then multiplied by a weight which is some kind of value corresponding to the sensitivity of the neuron. The weight is calculated by three ways perceptron rule in Equations 3.6 to 3.8 used when the data set is linearly separable, gradient descent in equations 3.9 to 3.11 when the data is not thresholded and not linearly separable but may lay in a local optimum value and sigmoid in Equations 3.12 and 3.13.

\[
\Delta w_i = w_i + \Delta w_i
\]

Equation 3.6 Weight values
In Equation 3.7 the weights adjusting rate is calculated by the target class value minus the predicted value multiplied by the value of the input attribute $x_i$ and the learning rate ‘n’. The predicted value is obtained by summing the current weight value $w_i$ multiplied by the input value $x_i$.

The sum of these values represents the activation which is then compared to a threshold value to determine if the sum is greater or equal to the threshold, the output is 1 if greater otherwise it is 0.

\[ a = \sum x_i \times w_i, \quad \hat{y} = \{a \geq 0\} \]

Equation 3.9 Activation function

\[ E(w) = \frac{1}{2} \sum (y - a)^2 \]

Equation 3.10 Error metric on weight w

\[ \Delta w_i = n(y - a) x_i \]

Equation 3.11 Weight change rate for gradient descent

The second way is gradient descent. In Equation 3.9 activation value is calculated. After that an error metric is calculated over the weight value defined as one half multiplied by the sum of the square of the target value ‘y’ minus the activation value ‘a’ previously calculated. This calculation is done for all the data set.

The weights for the input vector is adjusted every iteration with Equation 3.11 where ‘n’ is the learning rate, ‘y’ is the target value, ‘a’ the activation value without the threshold as it was for perceptron rule, where instead of ‘a’ was $\hat{y}$ which was $\{0, 1\}$, and $x_i$ the input value.

\[ \sigma(a) = \frac{1}{1 + e^{-a}} \]

Equation 3.12 Sigmoid equation for activation value 'a
And the last way is the sigmoidal function which draws a smoother curve as seen in Equation 3.13 as the activation ‘a’ grows lower the sigmoid tends to 0 otherwise it tends to 1.

As mentioned all units are located in layers as seen in Figure 3.2. The layers are organized as follows:

- **Input layer:** The input layer is the first layer in the structure and is in charge of receiving and arranging the inputs to be processed by the next layer(s) according to the mentioned equations above. The number of units depends on the number of attributes in the input data \{x_1, x_2, \ldots, x_n\}.

- **Hidden layer:** The hidden layer(s) receive the information from the input layer. In a hidden layer the information is processed according to the weighted input from the previous layer and outputs information through activation functions. Activation functions are thresholds that set limits to decide if the signal or values are enough to be sent to the next layer.

  Typically all the hidden units are interconnected giving the attribute of parallel processing characteristic of neural networks.

- **Output layer:** Finally information is sent to the output layer by the last hidden layer. In this layer information is summarized and output.

![Figure 3.2 Multilayer perceptron structure](image-url)
3.5 Reliability

As previously mentioned the goal of this research is to determine if feature selection can improve the results of machine learning algorithms. The data set has been used in previous work using area under the curve (AUC) and ALIFT indexes for performance measurement. In this report the following values are going to be used to provide reliability to the results and the comparison of the tested algorithms:

- Correctly and incorrectly classified instances: These are the indexes to measure accuracy of every classifier. It shows how many of the total instances were misclassified and how many were correctly classified, however these values could be compromised as the data set is unbalanced, which means there are more instances of one class than the other and this could affect the results.

- Kappa statistic with confusion matrix: Kappa measures how the classifier performed over the data set, this means the classification results against the true class of every instance however the interpretation of the value varies on context. It is possible to have a high kappa value but the classifier went wrong, for example the classifier correctly classified all ‘no’ class but only 2% of ‘yes’ class so the kappa is high but the model is useless since it misclassifies the class the bank wants to know. Kappa complements the other two used indices in that it considers possible classifications by chance, this happens when the classifier cannot decide between the classes offering a different perspective.

- ROC area: Shows for all classifications how well the model performed, which means the classifier adjusted correctly the thresholds for separating the classes according to the attributes. ROC curve overpasses accuracy index here, because even if the data set is unbalanced the curve shows if the classifier found the correct limits between classes with true and false positive rates. True positive rate is how often the classifier predicts positive (or yes) when the real classification is positive and false positive rate is how often the classifier predicts positive when the real classification is negative (or no). On the other hand if the data set contains 90% of one class even with a chance classifier would get 90% accuracy, which is not necessarily good performance.
3.6 Ethical considerations

Due to confidentiality issues the data set has been trimmed previously to this experiments for privacy purposes mentioned in [7].

4 Results/Empirical data

In this chapter the results collected from the experiments and the benchmarks of comparison between the algorithms and feature selection methods are presented.

4.1 Approaches for reporting results

Table 4.1 shows how the data set is composed:

<table>
<thead>
<tr>
<th>Class</th>
<th># of instances</th>
<th>% of total instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>4640</td>
<td>11.2654</td>
</tr>
<tr>
<td>No</td>
<td>36548</td>
<td>88.7345</td>
</tr>
</tbody>
</table>

Table 4.1 Data set composition

As can be seen in Table 4.1 the data set is unbalanced and to improve the results it is recommended to test different weights for the predicted class or using a method for unbalanced data sets as mentioned in Section 3.2.1.

For the experiments 10-fold cross-validation was used to divide the data set into training and testing subsets.

J48 is being used with default configuration by WEKA with a confidence factor of 0.25 and 2 minimum objects per leaf.

For MLP configurations the default learning rate is set to 0.3 by WEKA with 1 hidden layer and 500 epoch iterations for training.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly classified %</th>
<th>Incorrectly Classified %</th>
<th>Kappa Statistic</th>
<th>ROC Area</th>
<th>Time to build model (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>91.1892</td>
<td>8.8108</td>
<td>.5328</td>
<td>.884</td>
<td>7.92</td>
</tr>
<tr>
<td>J48/CfsSubsetEval</td>
<td>91.2669</td>
<td>8.7331</td>
<td>.5141</td>
<td>.910</td>
<td>/</td>
</tr>
<tr>
<td>J48/ReliefAttributeEval</td>
<td>91.1649</td>
<td>8.8351</td>
<td>.5319</td>
<td>.883</td>
<td>/</td>
</tr>
<tr>
<td>J48/Chi-squared attribute evaluation</td>
<td>91.483</td>
<td>8.517</td>
<td>.5455</td>
<td>.915</td>
<td>/</td>
</tr>
<tr>
<td>MLP (1 hidden)</td>
<td>89.7567</td>
<td>10.243</td>
<td>.4515</td>
<td>.886</td>
<td>270</td>
</tr>
<tr>
<td>layer)</td>
<td>3</td>
<td>9.3717</td>
<td>.4864</td>
<td>.921</td>
<td>571</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>MLP (10 hidden layers)</td>
<td>90.6283</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP (20 hidden layers and learning rate of 0.2)</td>
<td>90.2811</td>
<td>9.7189</td>
<td>.4865</td>
<td>.908</td>
<td>1769</td>
</tr>
<tr>
<td>MLP/CfsSubsetEval</td>
<td>90.7886</td>
<td>9.2114</td>
<td>.4728</td>
<td>.919</td>
<td>/</td>
</tr>
<tr>
<td>MLP/ReliefAttributeEval</td>
<td>89.9631</td>
<td>10.036</td>
<td>.4608</td>
<td>.885</td>
<td>/</td>
</tr>
<tr>
<td>MLP/Chi-squared attribute evaluation</td>
<td>89.8393</td>
<td>10.160</td>
<td>.468</td>
<td>.880</td>
<td>/</td>
</tr>
</tbody>
</table>

| Table 4.2 Prediction results |

- Correctly classified: correctly predicted test instances.
- Incorrectly classified: incorrectly predicted test instances.
- Kappa statistic: chance-corrected measure of agreement between the classifications and the true classes, > 0 means better than chance.
- ROC area: Receiver Operating Characteristic area is the space plotted between the true positive in function of the false positive rates [11].

### 4.2 Results Analysis

The results show that even the algorithms without feature selection offer acceptable results with over 80% accuracy. Feature selection and pre-processing methods can make a difference, although in this experiments it is not that large. For bigger data sets it can mean a big amount of money or clients. Table 4.2 shows that feature selection improved the obtained results in both accuracy and precision.

In the experiment 1% in MLP (one hidden layer vs CfsSubsetEval) means that over 400 possible clients are possibly saying yes, although it is not possible to be 100% sure to obtain those clients.

The percentages of correctly and incorrectly classified instances gives an idea of the performance of the algorithms. The results could however be affected by the unbalanced classes affecting the results.

Kappa statistic is used to measure the level of agreement between different observers which in this case is the algorithm over different instances where it finds the same results based on the analyzed instance values. The closer it gets to 1 the better and 0 means that the results are as accurate as pure chance [12]. Finally ROC area which is the area limited by the true positive rate (sensitivity) in function of false positive rate (specificity) that dictates
precision and exactitude of the generated model according to the analyzed instances [11]. Figure 4.1 shows how different ROC curves look and the meaning of the different possible forms for a ROC curve.

Moreover experiments for several configurations in hidden layers of MLP model were ran to observe different scenarios. This to compare a broader range of results where more resources (hidden layers) were assigned compared to attribute selection and test results in accuracy, precision and performance as seen in Table 4.2.

![Comparing ROC Curves](image1.png)

Figure 4.1 ROC curves comparison example.

![ROC curves comparison example](image2.png)

Figure 4.2 From left to right, ROC curve for J48 and J48 with correlation-subset feature selection
The important fact that ROC curves offer is the comparison of how the model tests different thresholds along the experiment, optimizing the results to get better positive classifications in this case. The thresholds are set to divide the data set to find the best measure and divide the instances for each class. In this case, according to all attributes the classifier decides where in the values the classes are best separated to offer the purest positive or negative, depending on the graph settings, division.

In Figures 4.2 and 4.3 it can be seen information about the behavior of the model, where the y-axis is the true positive rate and x-axis is false positive rate. This means the graph is oriented to determine how well the classifier did for the positive class and the closer the curve grows to the y-axis the better the model develops with more true positive classifications.

The colors of the graph shows how the performance of the model was affected by feature selection. As can be seen more clearly in the MLP graph, according to the proportions of classes in Table 4.1 as it goes closer to the y-axis, it keeps the orange color (higher true positive rate).

As for the improvements made by feature selection one needs to consider the extra processing time and resources needed to do this since the time taken for the combination MLP/Chi-Squared was about of two hours and considerable computational resources. There are many variables to be considered when selecting a data mining approach as the expected results depends on the information, the data set distribution and the available computational resources so at the end the user should consider if it is worth the extra improvements when using feature selection. Table 4.3 shows the total processing time for J48 and the different feature selection models.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Feature selection processing time</th>
<th>Classification time</th>
<th>Evaluation time</th>
<th>Total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>-</td>
<td>27s.</td>
<td>21s.</td>
<td>48s.</td>
</tr>
<tr>
<td>J48/CfsSubset Eval</td>
<td>1s.</td>
<td>5s.</td>
<td>4s.</td>
<td>10s.</td>
</tr>
<tr>
<td>J48/RELIEF</td>
<td>34:30m.</td>
<td>15s.</td>
<td>12s.</td>
<td>34:57m.</td>
</tr>
<tr>
<td>J48/Chi-squared</td>
<td>8s.</td>
<td>26s.</td>
<td>19s.</td>
<td>53s.</td>
</tr>
</tbody>
</table>

Table 4.3 Processing times.

5 Results Conclusion

A brief conclusion of results over both classifiers as well as an explanation of the results is presented in this chapter.

5.1 Decision Tree (J48 on WEKA or C4.5)

By analyzing the experiments for this classifier it is clear how feature selection affects the results, this does not mean it is better or worse but it rather depends on what results are expected and what results were obtained. In Table 4.2 the results show that feature selection improved accuracy as well as ROC area indices, which means the classifier was able to find a better threshold value to separate classes and improve classification accuracy than without feature selection. These results can be seen in the graphics as well (figure 4.2) where the curve grows from a higher point or value in the y-axis with feature selection and this is where the threshold shows that the classifier found better values to split the two classes (‘yes’ and ‘no’) which in this case is oriented to positive classifications.

5.2 Multilayer perceptron

For this classifier it is clear that feature selection improved results for both accuracy and ROC area indices, showing that the model performed better classifying instances and splitting the classes which in turn improves accuracy. It can also be seen on the graphs (see Figure 4.3) how the curve grows faster along the x-axis without feature selection even though both graphs start at the same point (0, 0). This point means the classifier is setting a threshold value where true positive rate is 0, which means the model is “supposing” all instances are negative, which is obviously not true.
On the other hand talking about performance one can set more resources to the MLP model to perform better with a trade-off over time.

In Table 4.2, MLP increases processing time, when adding hidden layers, as well as accuracy and ROC area finding better results around 10 hidden layers.

6 Discussion

In this section the results and how these are helpful to the target groups and the community in general are reviewed as well as a brief comparison with previous research and how these results complement or extend that research.

6.1 Problem solving/result

In [7] tests were ran using decision trees, neural networks and support vector machines from which the authors concluded that neural networks were the most precise and trustable using AUC and ALIFT indexes for reliability and performance measuring, at the same time semiautomatic feature selection by banking experience was applied and from 150 attributes 22 were selected. These 22 attributes were then used in the experiments carried out in this project.

The experiments in this project show that reliable results were obtained from the two used algorithms with 90% of correct classifications in almost all the cases showing how accurate the classifiers performed answering RQ2. The three used feature selection algorithms correlation-based feature selection, RELIEF attribute evaluation and chi-squared attribute evaluation also increased the results, with better overall results on chi-squared and J48 which benefits the most from feature selection methods for RQ3. Regarding RQ1 Table 4.2 shows that feature selection actually improved machine learning algorithms in accuracy and precision, as for performance Table 4.3 shows performance got improved as well in some cases.

6.2 Method reflection

Table 4.1 shows that the proportions of instances of each class in the data set where only 11% for ‘yes’ class and the rest for ‘no’ class meaning that the data set is unbalanced and this fact could affect the behavior of machine learning models if not dealt with. This is therefore an important point of improvement for future research. The probabilities that a client obtains a
service or a product offered by phone is truly low this is why obtained data will never be naturally balanced like in the data set used in this tests, but as mentioned in previous chapters there are methods to fix this, for example weighting the minority class or trimming the data set. There are many feature selection and pre-processing methods that could offer better results in performance, behavior, stability, precision and accuracy in machine learning algorithms. This is why research and testing different technologies is important, which at the end is the essence of engineering endeavors. Besides feature selection and data balancing there are other alternatives regarding telemarketing and sales since there are always new methods and/or ways to sell services and/or products from any field that could be less invasive and more efficient than having personnel calling and bothering people at their homes. This could also be translated to less costs but as good and important as telemarketing.

7 Conclusion

In this chapter the conclusion and possible research for the future is presented.

7.1 Conclusions

To conclude the goal of this thesis must be mentioned, which is to compare different feature selection methods with machine learning algorithms, focusing over feature selection effects on the classifiers and their performance in classifying tasks. Table 4.2 shows that feature selection methods improved results in precision and accuracy of the two classifiers giving around 1% better accuracy and around .1 better kappa values, which means the observed classifications improved in terms of the expected accuracy. The research questions have been answered with quantitative results and experiments. Reliability has been proved with measurable indices and objective interpretation (as objective as it can be) but it is worth mentioning that interpretation highly depends on expectations and desired results from the user. One could want to know the statistics of ‘negative’ class and decrease this values which not necessarily means to increase ‘positive’ classifications. This is where kappa and ROC areas help as they show the statistics of classifications, what increases or decreases true and false positive rates, finding class thresholds to find the limits of each class in terms of the attributes or comparing different results when using the same classifier as kappa does. At the end statistics importance is given by the user
as for this thesis, results were significantly improved by feature selection and can be concluded that feature selection is an important and worthwhile process in data mining.

7.2 Further Research

In the future it would be important to keep more information of the person’s mood in the last contact or its reaction to determine if it is feasible to call again or not, as well as to give different weights to the predicted classes and balance the data set some way presented here to see if results are close to current research or improved. Feature selection has proven to improve results in accuracy and performance of classifying, however these algorithms do not consider the state of the data set if it for example is unbalanced which could mislead the feature selection cutting the data set’s attributes because those selected determine one class but more information is needed for other classes. Another remark is to set expected values to determine the importance of the obtained results and the goal of the experiments to determine the information one wants to know and/or obtain from data mining processes.
References


Appendices