# THE USE OF ATTRIBUTE SELECTION IN THE BANKING SECTOR IN ORDER TO OBTAIN KNOWLEDGE OF CUSTOMERS

Dayana Carla de Macedo (UTFPR/PR) dayanamacedo@yahoo.com.br
Simone Nasser Matos (UTFPR/PR) snasser@utfpr.edu.br
Helyane Bronoski Borges (UTFPR/PR) helyane@utfpr.edu.br
Antônio Carlos de Francisco (UTFPR/PR) acfrancisco@utfpr.edu.br

**[1]Abstract**: The knowledge of consumers is important in order to trace the profile of customers. But to gain knowledge of customers as well as the characteristics is necessary for organizations to get the most from these data. Information related to the Customers at companies are collected and stored in databases. The administration of these data often requires the use of a computational tool. However the amount of data can be harmful to analysis the data. In the data mining context there is one task that is called of Attribute Selection. The Attribute Selection is a dimensionality reduction method in order to remove redundant information and improve the performance of the learning process as the speed as in the performance of classifier. The present search applied this method in order to comparison the Filter and Wrapper approach in the Customer domain, using one base in the customer domain, in the banking sector.

**Keywords:** Dimensionality Reduction. Attribute Selection. Customer Domain. Attributes.

**Resumo**: O conhecimento de consumidores é importante para traçar o perfil de cliente. Mas, para obter conhecimento de clientes, bem como, as características é necessário para as organizações a obtenção dessas a partir dos dados. Informações relacionadas aos clientes nas empresas são coletadas e armazenadas em banco de dados. A administração desses dados frequentemente demanda o uso de uma ferramenta computacional. Porém, a quantidade de dados pode ser prejudicial para a análise de dados. No contexto de mineração de dados existe uma tarefa chamada Seleção de Atributos. A Seleção de Atributos é um método de redução de dimensionalidade para remover a informação redundante e melhorar o desempenho do processo de aprendizagem e velocidade no desempenho do algoritmo classificador. A presente pesquisa aplicou este método para comparação da Abordagem Filtro e *Wrapper* no domínio de clientes, usando uma base do setor bancário.

Palavras-chaves: Redução de Dimensionalidade, Seleção de Atributo. Domínio Cliente. Atributos.

## 1. INTRODUCTION

Currently there is a need to comprehend new expectations, abilities and behaviors of final consumers, constantly shifting, allowing the identification of business opportunities. With the obtained insights, companies can identify innovative opportunities while reducing market risks. This kind of analysis favors the company through the fact that the knowledge obtained about their consumers will result in a reduced market risk at the time of the choice between which innovation project to be implemented.

The key to success for companies is in the creation of value through new products and services (innovations) shaped in the insights obtained from transactional data. The new competitive panorama requires the continuous analysis of data in search of insights, so as to recognize and interpret the emergent market tendencies.

It is the manager's job to provide mechanisms for the process of comprehending this information and extracting from new ideas from it. There is a need to use quantitative models in data analysis, in order to transform them into a form of knowledge useful in decision making (PRAHALAD, KRISHNAN, 2008).

Administrating and cherishing the information and data collected about clients requires the adherence of an information technology tool that contributes with the administration of information and the obtainment of knew knowledge.

In order for a continuous analysis to be done, there is a need to appeal to computational tools. In this way, many Data Mining techniques have been employed, using specific algorithms for the extraction of patterns. These patterns allow the obtainment of potentially useful information (CHEN; HEN; YU, 1996; MITRA; PAL; MITRA, 2002).

There are successful Data Mining applications in several areas including WEB, marketing, financial and banking, telecommunications, among others (ROMDHANE; FADHEL; AYEB, 2010).

However the amount of data can be harmful to analysis the data. In the data mining context there is one task that is called of Attribute Selection that can assist this process. The present search applied this method in order to comparison the Filter and Wrapper approach in the Customer domain, using one base in the customer domain, in the banking sector.

This article is organized as follows: The section 2 shows the literature about dimensionality reduction. The section 3 reports the current situation regarding the database. The section 4 describes the methodology used in this paper. The section 5 reports the results found using the Filter and Wrapper Approach. After, the section 6 shows the applicability of the results. Finally, the last section presents the final considerations of this work.


## 2. DIMENSIONALITY REDUCTION

As far as the dimensionality, theoretically, it is intuitive to think that the higher the amount of attributes, more information would be available supposedly for the algorithm of data mining. Nevertheless, with the growth of attributes in the data, these tend to become sparser (DY, 2007).

When it comes to dimensionality reduction, two approaches are commonly used: attributes selection and attributes transformation. According to Guyon and Elisseeff (2003) the attributes selection has as a goal the elimination of redundant and irrelevant attributes. The process of attributes selection refers to selection of an attributes subset from the original database, through of following of certain criteria. Thus, the selected attributes keep the original physical interpretation, making easier the comprehension of the generated model (LIU et al, 2005).

However, the transformation of attributes may bring several benefits such as: facilitation of the comprehension, visualization of the data and the reduction of the computational cost of the algorithm of data mining applied.

The process of dimensionality reduction is an important task in the process of acquiring automatically knowledge to determine the best variables for modeling, but also for comprehensibility and scalability aspects of the resulting models (KIM et al., 2000).

The process of attributes selection ensures the quality that the data get to the mining phase (LIU; SETIONO, 1996).

Whether the target task is the classification, the attributes selection will search to minimize the error rate of the classifier, the complexity of the knowledge to be generated for it, and the number of selected attributes to compound the new base (BORGES; NIEVOLA, 2007).

According to Dash and Liu (1997) the method of attributes selection consists of 4 (four) main steps as the figure 1 shows.
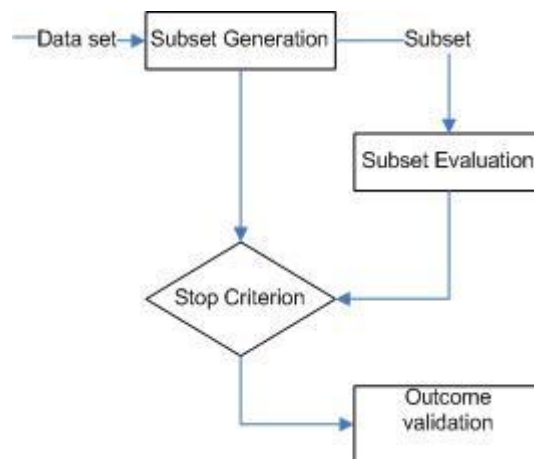


**Figure 1 - Basic steps of the attributes selection process (Dash and Liu; 1997, p.133)**

Thus, it is noted that the process input is given for a data set of one only base. Therefore, it is described in the next section the method proposed in which it is analyzed more than one base in one domain.


## 3. CURRENT SITUATION REGARDING THE DATABASE


Currently, most databases follow the relational model and relational object. The relational model is based in the presentation of data in the form of tables. Each column of a table has a unique name, and one line of the object represents a relationship between a group of values. The relational model object allows for relations outside of the first normal form and other characteristics of an object-oriented model (SILBERSCHATZ, KORTH, SUDARSHAN, 2006).

The development of models, also called profiles, is an important step for directed marketing. Many marketing managers can develop pleasant long-term relationships with their consumers if they detect and predict changes in their consuming habits and behavior (ROMDHANE; FADHEL; AYEB, 2010).

When concerning the customer, CRM is the tool of choice, consisting of four dimensions that are customer identification, customer attraction, customer retention and customer development. Many data mining tasks are applied alongside CRM to develop these four dimensions (NGAI; XIU; CHAU, 2009).

In this way, many Data Mining techniques have been employed using specific algorithms for the extraction of patterns. These patterns allow the obtainment of previously unknown knowledge and potentially useful information (CHEN; HEN; YU, 1996; MITRA; PAL; MITRA, 2002).

Recently, the segmentation of customers is based in transactional and behavioral data, such as: types, volume, shopping history, call center complaints, vindications, WEB access data (LEE; PARK, 2005). From these data it is possible to obtain the patterns that contribute to the generation of useful knowledge.

Decision trees are employed to extract models and describe sequences of interrelated decisions or predict future tendencies among customer data (BERRY; LINOFF, 2004; CHEN et al., 2003; KIM et., 2005).

There is also the use of rules of association with the discovery of potential relations between data, which permits the construction of models to predict the value of a future customer (WANG et al., 2005).

Many Data Mining techniques are applied to the development of each of CRM's dimensions. In customer retention, techniques like Clustering, Sequence Discovery, Classification and Association Rules are used focusing on loyalty programs and customer complaint, which are elements of CRM.

With relation to the customer development dimension, the tools of task classification, clustering, regression, association and sequence discovery are also employed. These techniques are applied in CRM elements such as: cost of living and Market Basket Analysis, which analyses the customer's behavior through consuming habits.

There are successful Data Mining applications in several areas including WEB, Marketing, financial and banking, telecommunications, among others. Romdhane, Fadhel and Ayeb (2010) have developed an approach for customer profiling composed of three stages, the first one consisted in using the FCM-based algorithm to group customers, also using information entropy[2]. The last stage consisted in building a customer profile through a neural network called backpropagation (ROMDHANE; FADEL; AYEB, 2010). The research resulted in the prediction of consuming habits and customer behavior, the organizations are able to develop long-term relationships with these customers.

In the work of Park and Chang (2009), there was a development of a model of a customer profile based on group and individual behavior information such as clicks, shopping cart insertions, purchases and fields of interest, also using Data Mining.

Jiang and Tuzhilin (2009) point out, in their work, that the quantity of available data for the creation of models is one of the biggest difficulties encountered by Data Mining, that is, the accommodation, the irregularity in the data and the necessity to capture the imprecise nature of human behavior.

A new approach of hybrid algorithm, called Hybrid Evolutionary Algorithm in the selection of attributes, with the goal of reducing dimensionality. By means of the obtained results, it was possible to conclude that the use of this algorithm produced good results in relation to the classifiers and a high level of consistency when compared to other algorithms (TAN; TEOCH; GOH, 2009).

The works in the area of Attribute Selection seek the reduction of dimensionality, but there was no work found using the algorithms CFS, CSE in Filter approach and Naïve Bayes, J48

---

[2]Information entropy is used to quantify the importance of an attribute.

and SVM in the Wrapper approach for the domain of customers. Thus, this search used one database in the banking sector.


# 4. METHODOLOGY


In this step, it was made the database selection, that is, the sample. The amount of the sample to this work was 1 (one) database. The experiment in this research focuses on the customer profile; the base selected was in this domain. In order to realization of the experiment, database of public domain was selected. Thus is described the database that was used in this work, as follows (UCI, 2011).

- Database 1 – Stalog: credit data of a German bank, that is, in the banking sector: the base contains data about people that classified them according to the risk of credit that these people may present; being classified as good or bad. The base has a total of 1000 instances, with 20 attributes, where 7 are numerical and 13 are categorical.

This work was operationalized into five steps: Choice of Database, Preparation of Database, Application of the Attribute Selection, Execution of Classification Algorithms and Evaluation of the obtained results, as illustrated in Figure 2.
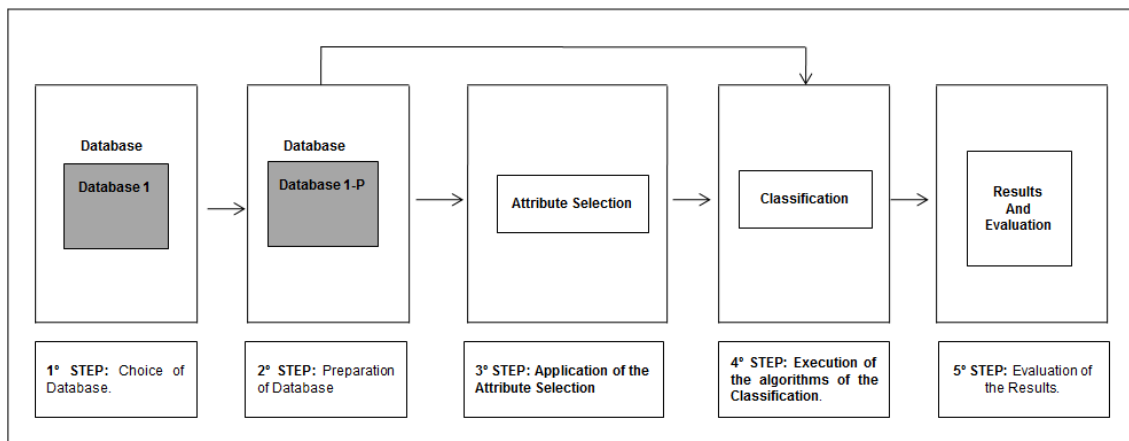


**Figure 2 - General process of work development**

**Source: Author**

The choice of Database this stage comprises the selection of the database, that is, the sample. As the experiment in this research deals in client, the selected base were in the Client domain, which are available for download from the Machine Learning Repository (UCI, 2011) website. For the realization of the experiment, database from public domain which contained client data and attributes were selected.

The database preparation stage involves some processes, which are: Identifying the environment in order to Data Mining; Performing Database and preparing the database to the environment to the Data Mining. After, the Application of Attribute Selection consists of the submission of database to the Attribute Selection Method applying Filter and Wrapper.

The Frame 1 illustrates the search criteria and the evaluation measure used in order to create each subset in the database prepared in the Filter and Wrapper Approach.

| | Algorithm | Search Criteria | Evaluation Measure | Subset |
|---|---|---|---|---|
| Filter | CFS | Sequential | Dependence | Subset 1.1 |
| Filter | CSE | Sequential | Consistency | Subset 1.2 |
| Wrapper | Naïve Bayes | Sequential | Wrapper (Naïve Bayes) | Subset 1.3 |
| Wrapper | J48 | Sequential | Wrapper (J48) | Subset 1.4 |
| Wrapper | SVM | Sequential | Wrapper (SVM) | Subset 1.5 |

**Frame 1 – Algorithms of Attribute Selection**

**Source: Authors**

The subset of attributes more relevant created are characterized as the output of Attribute Selection Algorithm. In the present search was created the amount of 5 subsets for each database, Stalog as showed in the Frame 1.

The classification step of the experiment consists in the execution of classification algorithms and the WEKA tool. The entrance of the classification algorithm consists in the obtainment of the base from the second stage as well as the base and subsets originated in the third stage. The classification algorithms used in the search are Naïve Bayes, J48 and SVM. The fifth stage of this research is destined to evaluate the method of attribute selection. This stage comprehends the phases of post-processing of the KDD process.

The next section describes the each step with details in order to understand the present experiment.

## 4. 1 Choice of Database

This stage comprises the selection of the database, that is, the sample. The sampled quantity for this work was of 1 (one) base. One example application, in this work, represents a segment of a given domain. Considering Data Mining, there is no established minimum.

As the experiment in this research deals in client, the selected base was in the Client domain, which is available for download from the Machine Learning Repository (UCI, 2011) website. For the realization of the experiment, database from public domain, which contained client data and attributes, was selected. The next section reports the step called database preparation.

## 4. 2  Preparation of Database

The database preparation stage involves some processes, which have been described by the algorithm illustrated in Frame 2.

**Algorithm:** Preparing Database
**Input**: Database 1
**Process**:
- Identifying the environment in order to Data Mining
- Performing Database
- Preparing the database to the environment to the Data Mining

**Output**: Database 1-P

**Frame 2 – Preparing Database Algorithm**

**Source: Author**

The Preparing Database Algorithm has, as entry, the base that were selected in the first step, that is, Database 1, and after the execution of the tree processes, the output are the prepared base, containing consistent data. The processes are described as follows.

## 4.2. 1 Identifying the environment for Data Mining

In order to apply Data Mining, it is necessary to define the environment for the execution of its tasks. This work utilized WEKA, that is a public domain package consisting of algorithm implementations and several Data Mining techniques, written in Java, with the perk of being portable, that is, it can be executable in many platforms (WEKA, 2010).

With the environment set for Data Mining, there is the need to prepare the base in the format that the tool of choice, in this case, WEKA, executes its algorithms. Thus, it is necessary to clean the database. The next section presents this process.

## 4.2.2 Performing data cleaning

This process comprehends the pre-processing phase of the KDD process, that is characterized as the cleaning and treatment of data, aiming for integrity.

At the end of the experiment's first stage, the selected base was downloaded. The base was in text format (.txt). The .txt format file was opened in a spreadsheet application, in this case Excel, and each base was transformed in a .xls file format for the standardization of data cleaning. The sub-steps correspond to the steps belonging to the pre-processing of the Knowledge Discovery in Database (KDD) process, namely the cleaning, codification and enrichment of data.

After the sub-steps, the database was formatted in .arff files, as required by WEKA. The next section reports the last process of this step.

### 4.2.3 Preparing the base for the Data Mining environment

For the Data Mining environment, the database must be in the adequate file format. Following the creation of .arff (WEKA format) files, after cleaning, codification and enrichment, the spreadsheet data was saved initially in the Comma-Separated Values extension (CSV).

Upon saving the file in this format, the values belonging to the attribute domain are divided only by semicolons (;), but WEKA needs them to be divided only by commas (,). This way, the .csv files were opened in an application capable of reading text, like Notepad, and the semicolons were replaced by commas.

After this procedure, the file was saved in the .arff extension. As a result, the prepared base was obtained, named: Base 1-P.

### 4.3 Application of Attribute Selection

This section aims to describe the third step of the search where is applied the Attribute Selection.

The database prepared, namely, in this experiment, Stalog (Database 1-P) is submitted to the Attribute Selection Method. This method will be detailed by the algorithm called Attribute Selection, illustrated in the Frame 3.

**Algorithm**: Attribute Selection
**Input**: Database 1-P
**Process**:
  • Applying the Filter and Wrapper Approach
**Output**: Selected Attributes
        Subset 1.1 Database 1-P
        Subset 1.2 Database 1-P

**Frame 3 – Algorithms of Attribute Selection**

**Source: Authors**

The input of this algorithm consists in the database which was prepared to the WEKA environment (Database 1-P). Thus, each one of the database was opened in the WEKA environment and submitted to the process of Attribute Selection algorithm. The processes of this algorithm involve applying the Filter and Wrapper approach described as follows.

### 4.3.1   Applying Filter and Wrapper

In the Filter Approach each one of the database was submitted to the algorithms Correlation Feature Selection (CFS) and Consistency Subset Eval (CSE). But, in the Wrapper approach were executed the algorithm: Naive Bayes, J48 and SVM in each database.

After the choice of the algorithms, it is defined which search method and evaluation criteria must used to the execution of the selection algorithms. Thus, when these algorithms are used these is created several attribute subsets named of: Subset 1.1 Database 1-P, Subset 1.$_a$

Database 1-P, where "a" represents the amount of algorithms used. An attribute subset represents the attributes more relevant of a determined base, with their respective instances.

The Frame 4 illustrates the search criteria and the evaluation measure used in order to create each subset in the database prepared.

| | Algorithm | Search Criteria | Evaluation Measure | Subset |
|---|---|---|---|---|
| Filter | *CFS* | Sequential | Dependence | Subset 1.1 |
| | *CSE* | Sequential | Consistency | Subset 1.2 |
| Wrapper | *Naïve Bayes* | Sequential | *Wrapper* (*Naive Bayes*) | Subset 1.3 |
| | *J48* | Sequential | *Wrapper* (*J48*) | Subset 1.4 |
| | *SVM* | Sequential | *Wrapper* (*SVM*) | Subset 1.5 |

**Frame 4 – Algorithms of Attribute Selection**

**Source: Authors**

The subset of attributes more relevant created are characterized as the output of Attribute Selection Algorithm. In the present search was created the amount of 5 subsets for the database, Stalog.

After the subsets were submitted to execution of the classifiers algorithms, however, this procedure belongs to the fourth step of the experiment that will be detailed in the next section.

## 4.4 Execution of Classification Algorithms

The classification step of the experiment consists in the execution of classification algorithms and the WEKA tool. For this stage, refer to the classification algorithms in Frame 5, summarizing the adopted procedure.

**Algorithm**: Classification
**Input**: Database 1-P
       Subset 1.1, ...., Subset 1.$_a$ Database 1-P
**Process**: Application of the Classifiers Algorithms (*Naive Bayes*, *J48* e *SVM*)
**Output**: Hit Rate of the Classifiers

**Frame 5 - Classification algorithm**

**Source: Author**

The entrance of the classification algorithm consists in the obtainment of the base from the second stage as well as the base and subsets originated in the third stage. Following is a description of this algorithm.

### 5.4.1   Application of classification algorithms (*NaiveBayes*, *J48* and *SVM*)

This process occurs for the subsets of attributes generated in the attribute selection method as well as the Framework concepts method.

Employs the Cross-validation method for the execution of the classification algorithms. This method consists in subdividing the database in 3 parts, where 2 consist in a training base and the other part is destined for testing. In other words, a proportion of 2/3 of training data and 1/3 for testing. For all generated subsets, training and testing base were created.

After this division, initially two thirds of the base are destined to be a part of the Training base 1 (Part 1 and Part 2), and one third will be the Testing base (Part 3). For the posterior Training and Testing base, the procedure must be repeated, with the alternation of the Testing base. If in test 1 it was Part 3, in the upcoming tests it should be parts 2 and 1. The Figure 3 can be show this process of cross-validation:
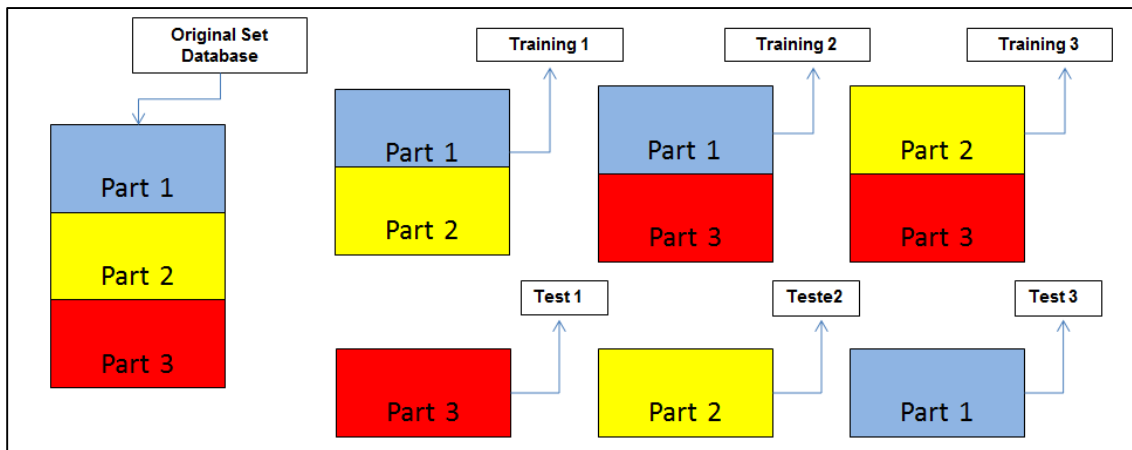


**Figure 3: Partitioning in training and tests – Cross-Validation Method**

**Source: Authors**

The next step is applying the classification algorithms: Naïve Bayes, J48 and SVM, that are already encoded in the WEKA environment. Finishing the execution of algorithms, the success rates are saved for input in the next stage.

## 4.4 Evaluation of Results

The fifth stage of this research is destined to evaluate the method of attribute selection. This stage comprehends the phases of post-processing of the KDD process.

The evaluation algorithm, shown in Frame 6, illustrates the steps that were taken during this last stage of development.

---

**Algorithm**: Evaluation
**Input**: Hit Rate of the Classifiers for each one of the following elements:
        Database 1-P
         Subset 1.1, ...., Subset 1.$_a$ Database 1-P
**Process**: Calculating Average and Standard Deviation
**Output:** Average Values
      Best and Worst Algortihms
      Relation of the Attributes using the Attribute Selection

---

**Frame 6 - Evaluation algorithm**

**Source: Author**

This algorithm utilized as input all the success rates obtained from the execution of classification algorithms, considering all base and subsets generated from the second to fourth stages.

In this algorithm, there are two main processes, which will be described below.

### 4.4.1 Calculating mean and standard deviation

Since Cross-validation of 2/3 was used, for each obtained database, we obtain three values of success rate, as per the generation of, always, three training base and three testing ones.

In this process, several evaluations are realized, like the calculation of the arithmetic mean and the standard deviation considering the success rates obtained in during the execution of the trainings and testing for each of the base in each of the classifications, with the characteristics:

- All attributes (Base 1-P);

After these evaluations, a general comparison between the best and worst performances of classification algorithms for each of the base was done.

The next section describes the results attained by means of the present experiment.

## 5. RESULTS OF ATTRIBUTE SELECTION ON THE DATABASE

In order to apply the attribute selection were used two approaches: Filter and Wrapper. It was used as the search criteria the sequential for both approaches. However, in the Filter, approach the evaluation measure or criteria used for the CFS algorithm was the dependence and for CSE algorithm was the consistence.

Table 1 shows the attribute numbers created in each subset for the three databases: Stalog, Customer and Insurance, as well as, for each algorithm:

Table 1 – Numbers of Attribute selected for each database

| Algorithms | Stalog |
|---|---|

| | | | |
|---|---|---|---|
| **Filter** | Subset 1.1 | CFS | 3 |
| | Subset 1.2 | CSE | 14 |
| **Wrapper** | Subset 1.3 | Naïve Bayes | 6 |
| | Subset 1.4 | J48 | 6 |
| | Subset 1.5 | SVM | 10 |

<center>**Source: Authors.**</center>

The Subset 1.1 has the subset of attributes selected in each database for the CFS algorithm, as well as, for the others subsets considering the CSE algorithms, J48, Naive Bayes, SVM respectively (Subset 1.2, …, Subset 1.5).

## 5.1 Filter Approach

As mentioned, the database Stalog was submitted to Filter Approach, for the CFS and CSE algorithm. With the execution of theses, algorithms were created two attribute subsets.

Thus, with the application of CFS algorithm was created the Subset 1.1 and for the CSE algorithm was originated the Subset 1.2. The Subset 1.1 consists of the attribute subset created in each database, Stalog, as well as, the subset of attributes 1.1 considering the CSE algorithm.

After the generation of attributes, the classification task with the J48, Naive Bayes and SVM algorithms using the cross-validation method. Table 2 illustrates the results of the average hit rate of the classifiers and standard deviations for the attribute subsets created with the CFS algorithm.

<center>**Table 2 – Results of the classifiers to the CFS Algorithm.**</center>

| | Subset 1.1 | | | | | |
|---|---|---|---|---|---|---|
| | Naïve Bayes | | J48 | | SVM | |
| | Average | D.S.[3] | Average | D.S. | Average | D.S. |
| Stalog | 73,50% | 1,39% | 71,60% | 1,66% | 70,80% | 0,32% |

<center>**Source: Authors.**</center>

It is possible to check that the Naïve Bayes had the best performance in Stalog database. But, the SVM algorithm was considered which had the worst performance for in this database. However, each database presented average result to the three algorithms, thus, there were not high dispersions.

Table 3 illustrates the average and standard deviations of results of the classifiers with respect to the Subset 1.2 created with the application of CSE algorithm.

<center>**Table 3 – Results of the classifiers to the CSE Algorithm**</center>

| | Subset 1.2 | | | | | |
|---|---|---|---|---|---|---|
| | Naive Bayes | | J48 | | SVM | |
| | Average | D.S. | Average | D.S. | Average | D.S. |

---

[3] D.S means Deviation Standard.

| | | | | | | |
|---|---|---|---|---|---|---|
| Stalog | 74,30% | 2,55% | 73,27% | 2,00% | 75,00% | 3,74% |

Therefore with the results is observed that in the Stalog database the algorithm which had the best performance was SVM. The worst performance of the algorithms in the Stalog database was checked with the J48 algorithm, with the average value 73,26%.

Through of the hit rate obtained by the Filter Approach is possible to affirm that the best performance was observed for the CSE algorithm presented a best performance only in the Stalog database.

The next section brings the obtained results to the Wrapper Approach.

## 5.2 Wrapper Approach

This approach consists of the application of the J48, Naive Bayes and SVM to generation of the attribute subsets. In this step were created three subsets using the attribute selection method. The subset 1.3 through of the application of the algorithm Naïve Bayes, the Subset 1.4 using the SVM algorithm and the Subset 1.5 through of J48 algorithm.

After generation of the subsets, these were submitted to the classification task, in which also was made the database partitioning in training and tests to the application through cross-validation method.

After database partitioning, the Subset 1.3 created was submitted to classification task, using only the Naive Bayes algorithm. In the subset 1.4 was submitted to the J48 algorithm and in the last subset 1.5 was executed the SVM algorithm. The Table 4 presents the results obtained to the Naive Bayes, J48 and SVM.

**Table 4 – Results of the Classifiers to the Wrapper Approach**

| | Subset 1.3 | | Subset 1.4 | | Subset 1.5 | |
|---|---|---|---|---|---|---|
| | Naive Bayes | | J48 | | SVM | |
| | Average | D.S. | Average | D.S. | Average | D.S. |
| Stalog | 73,70% | 3,81% | 73,20% | 0,33% | 74,40% | 1,90% |

The best performance of the classifier algorithm in the Stalog database was the SVM algorithm where it presented an average value of hit rate of 74,40%. Considering the Stalog database the worst method of the attribute selection was using the J48 algorithm, because it presented the lowest value of hit rate of this method.

It is possible to verify that the SVM algorithm presented the best performance among the other algorithms, showing that it was the best attribute selection method in the Wrapper Approach. Also with respect to the worst selection methods is possible to highlight the J48 in this approach.

## 5.3 Comparison of the Filter and Wrapper Approaches

Posteriorly to the results of the execution of algorithm of attribute selection, using the Filter and Wrapper Approach, these results were submitted to an analysis in order to verify which the best approach to the generation of subsets was.

Table 6 shows the average values of the classifiers obtained to the Subset 1.1 Stalog with all the attributes and the subsets submitted to Filter Approach (CFS and CSE) and Wrapper. It is important to highlight that in the Wrapper Approach the values illustrated are the results of the classification of each algorithm. However, in relation the analysis of the Filter Approach it is necessary the use of arithmetic average to point which of two algorithms, CFS and CSE, was the best among then. In the Wrapper Approach is done the isolated analysis of the results, without applying the arithmetic average obtained of each classifier to check of the Naïve Bayes and SVM algorithms, which was the best in the attribute selection.

**Table 5 - Results of the Classifiers to the Subset Stalog using the Attribute Selection Method**

| | Stalog | | | | | |
| | Naive Bayes | | J48 | | SVM | |
| | Average | D.S. | Average | D.S. | Average | D.S. |
|---|---|---|---|---|---|---|
| All Attributes | 57,12% | 0,76% | 54,32% | 2,37% | 56,61% | 1,14% |
| CFS | 73,50% | 1,39% | 71,60% | 1,66% | 70,80% | 0,32% |
| CSE | 74,30% | 2,54% | 73,27% | 2,00% | 75,00% | 3,74% |
| Wrapper | 73,70% | 3,81% | 73,20% | 0,33% | 74,40% | 1,90% |

Source: Authors.

It is observed that in this subset the best results were with the use of attribute selection in the Wrapper Approach, together with SVM algorithm. The CFS and CSE algorithms in the average of the classifiers were not higher to Wrapper. The worst results presented were observed without the use of the Selection Method.

## 6. APPLICABILITY OF THE RESULTS

From the identification of the best subsets it is possible to use the Data Mining tasks to identify the customer's profile.

According to the observed results, the best subsets of attributes of generated in each base were identified. For the *Stalog* base, the subset 1.5 showed the best results of success rate. With classification by the J48 algorithm and with the result of the use of the applicability algorithm, a set of rules was created, building a model for classifying instances.

The Figure 4 exemplifies one of the rules generated for the attribute called Status of existing checking account, which indicates the situation of the transactional account existing in the *Stalog* base.

```
Status of existing checking account = A11
|  Credit history = A30: 2 (13.0/3.0)
|  Credit history = A31: 2 (22.0/6.0)
|  Credit history = A32
|  |  Number of existing credits at this bank <= 1
|  |  |  Purpose = A40: 2 (41.0/15.0)
|  |  |  Purpose = A41: 1 (13.0/4.0)
|  |  |  Purpose = A42
|  |  |  |  Duration in month <= 16: 1 (16.0/2.0)
|  |  |  |  Duration in month > 16
|  |  |  |  |  Credit Amount <= 3518: 2 (17.0/5.0)
|  |  |  |  |  Credit Amount > 3518: 1 (11.0/3.0)
|  |  |  Purpose = A43
|  |  |  |  Duration in month <= 33: 1 (29.0/11.0)
|  |  |  |  Duration in month > 33: 2 (5.0)
|  |  |  Purpose = A44: 2 (5.0/1.0)
|  |  |  Purpose = A45: 2 (1.0)
|  |  |  Purpose = A46: 2 (7.0/2.0)
|  |  |  Purpose = A47: 1 (0.0)
|  |  |  Purpose = A48: 2 (1.0)
|  |  |  Purpose = A49
|  |  |  |  Duration in month <= 36: 1 (2.0)
|  |  |  |  Duration in month > 36: 2 (2.0)
|  |  |  Purpose = A410: 1 (2.0)
|  |  Number of existing credits at this bank > 1: 2 (8.0/2.0)
|  Credit history = A33
|  |  Duration in month <= 18: 1 (3.0)
|  |  Duration in month > 18: 2 (9.0)
|  Credit history = A34: 1 (67.0/18.0)
```

**Figure 4 - Rules generated with the employment of the J48 algorithm**

**Source: Author**

From the generated rules it is possible to come to conclusions in respect of the customer in the decision-making process. For example, according to the rule generated in Frame 14, the A11 value assigned to attribute 1 indicates the customers that currently have no financial activity in this bank. These customers have an effective credit operation, since because of the attribute 16 it is possible to identify that they possess only one operation. Out of the customers that already performed loans in this bank, the generated rule uses the values A30, A31 and A32, which indicate that they fulfilled their obligations within the stated period.

The class attribute of base *Stalog* aims to classify a customer as good or bad. This way, through application of algorithm J48, decisive rules and relations for decision-making about the customer can be acquired. Moreover, behavioral knowledge is also obtained.

In this case the relationships are established between attributes and their values. With the use of a determined classification algorithm, after the dimensionality reduction procedure the manager has the pattern and knowledge of their customers through data analysis. From this, market strategies are defined to maintain competitiveness. Through reduction algorithms, the quantity of generated rules is smaller, containing only the most relevant attributes, in order to facilitate data analysis.

# 7. CONCLUSION

The application of the dimensionality reduction methods is important, because the search of useful knowledge and standards in databases does not require the presence of a significant number of attributes.

The use of databases with all the algorithms can harm the performance of the learning process of the algorithms. Thus, there is the need of applying methods that ensure the data quality that arrive in the data mining step. The amount of redundant information can confuse the algorithm and does not assist in the search of a correct model to the knowledge.

These two methods were applied in the customer domain, using the banking segment. In this paper, the database was denominated: Stalog, respectively.

Analyzing the results obtained, using as criteria of evaluation the cross-validation verified that the used of these methods resulted in an improvement of the hit rate when is compared with the databases having all the attributes.

Among the methods of Attribute Selection, Filter and Wrapper Approach the best result found to Stalog was using the Wrapper Approach. However, the best approach the Wrapper, using the SVM algorithm. This approach normally presents a higher performance compared to other algorithms according to the literature, fact also observed in this paper.

Thus, the use of a determined classification algorithm, after the dimensionality reduction procedure the manager has the pattern and knowledge of their customers through data analysis.

# REFERENCES

BERRY, M. J. A.; LINOFF, G. S. **Data mining techniques:** for marketing, sales, and customer relationship management. Indianapolis, Ind.: Wiley, 2004.

BORGES, H. B.; NIEVOLA, J. C. Comparing the dimensionality reduction methods in gene expression databases. Expert Systems with Applications, v. 39, n 12, p. 10780-10795, set. 2012.

CHEN, M.S.; HEN, J.; YU, P. S. Data Mining: an overview from a database perspective. **IEEE Transactions on Knowledge Data Engineering,** v. 8, n. 6, p. 866-883,1996.

CHEN, Y. L., HSU, C. L., CHOU, S. C. Constructing a multi-valued and multilabeled decision tree. **Expert Systems with Applications**, v. 38, n. 2, part 1, p. 4339-4347, ago. 2003. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417403000472>. Acesso em: 20 de dez. 2011.

DASH, M.; LIU H. Feature selection for classification. **Intelligent Data Analysis - an International Journal,** v. 1, n. 3, p. 131-156, 1997.

DY, J. G. Unsupervised feature selection. In: COMPUTACIONAL methods of feature selection. London: Chapman & Hall/CRC, 2007. Cap. 2, p. 19-39.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research,** v. 3, p. 1157-1182, 2003.

JIANG, T.; TUZHILIN, A. Improving personalization solutions through optimal segmentation of customer bases. **IEEE Transactions on Knowledge and Data Engineering,** v. 21, n. 3, p. 305-320, mar.2009. Disponível em: <http://pages.stern.nyu.edu/~atuzhili/pdf/701_Jiang_Tianyi.pdf>. Acesso em: 19 jul. 2011.

KIM, J. K., SONG, H. S., KIM, T. S.; KIM, H. K. Detecting the change of customer behavior based on decision tree analysis. **Expert Systems with Applications**. v. 22, n. 4, part 1, p. 193-205, set. 2005. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0394.2005.00310.x/abstract>. Acesso em: 20 de dez. 2011.

LEE, S.; PARK, Y. Customization of technology roadmaps according to road mapping purposes: Overall process and detailed modules. **Technological Forecasting & Social Change,** v. 72, p. 567-583, 2005. Disponível em: <http://www.maoner.com/Cited_Saritas_Oner_2004.pdf> Acesso em: 5 jul. 2012.

LIU, H., YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on knowledge and Data Engineering.** v. 17, n. 4, p. p. 491-502, 2005. Disponível em: <http://www.public.asu.edu/~huanliu/papers/tkde05.pdf> Acesso em: 8 jul. 2011.

MITRA, S.; PAL, S. K.; MITRA, P. Data mining in soft computing Framework: a survey. **IEEE Transactions On Neural Networks,** v. 13, n. 1, p. 3-14, 2002. Disponível em: <http://repository.ias.ac.in/26054/1/310.pdf>. Acesso em: 6 ago. 2012.

NGAI, E, W, T.; XIU, L.; CHAU, D. C. K. Application of data mining techniques in customer relationship management: a literature review and classification. **Expert Systems with Application,** v. 36, n. 2, part 1, p. 2592-2602, mar. 2009.

PARK, Y. J.; C

HANG, K. N. Individual and group behavior-based customer profile model for personalized product recommendation. **Expert Systems with Application,** v. 36, n. 2, part 1, p. 1932-1939, mar. 2009. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417407006288> Acesso em: 8 jul. 2011.

PRAHALAD, C. K.; KRISHNAN, M. S.The New Age of Innovation: Driving Cocreated Value Through Global Networks. McGraw-Hill Professional, New York; 1 edition, 2008.

ROMDHANE, L. B.; FADHEL, N.; AYEB, B. An efficient approach for building customer profiles from business data. **Expert Systems with Applications,** v. 37, n. 2, p. 1573-1585, mar. 2010.

SILBERSCHATZ, A.; KORTH.; H. F.; SUDARSHAN, S. Sistemas de banco de dados. 5. Ed. Tradução de Daniel Vieira. Rio de Janiero: Editora Elsevier, 2006.

TAN, K. C.; TEOH, E. J.; YU, Q., GOH, K. C**.** A hybrid evolutionary for attribute selection in data mining**. Expert Systems with Application,** v. 36, n. 4, part 1, p. 8616-8630, mai. 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095741740800729X>. Acesso em: 20 dez. 2011.

UCI Machine Learning Repository. **Browse Through:** 5 data sets. Disponível em: <http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=bus&numAtt=&numIns=&type=&sort=nameUp&view=table>http://archive.ics.uci.edu/ml/>. Acesso em: 20 dez. 2011.

WANG, K., ZHOU, S., YANG, Q.; EUNG, J. M. S. Mining customer value: from

association rules to direct marketing. **Data Mining and Knowledge Discovery,** v. 11, p. 57-79, 2005.

WEKA. The University of Waikato. **Software.** Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 24 jan. 2011.