Solutions for analyzing CRM systems - data mining algorithms

ADELA TUDOR, ADELA BARA, IULIANA BOTHA
The Bucharest Academy of Economic Studies
Bucharest
ROMANIA
{adela_lungu}@yahoo.com
{Bara.Adela, Iuliana.Botha}@ie.ase.ro

Abstract: - The main goal of the paper is to illustrate the importance of the optimization methods used in data mining process, as well as the specific predictive models and how they work in this field. Also, the customer relationship management systems have been developed lately, offering new opportunities for a strong profitable relation between a business and clients.

***Key words:* -** data mining, data warehouses, attribute selection, optimization algorithms, preprocessing, clustering, classification, association, customer relationship management, OLAP

1. Introduction

All over the world, we are surrounded by data. There are large volumes of data, but not enough information. This is a problem that many companies and industries are facing. A solution to this issue could be data mining, also known as "knowledge discovery in databases". Many institutions hold large amounts of data, due to their work activity and also due to the rapidly evolution of technology. Useful information is hiding behind these data, but traditional methods of data analysis are outweighed by large volumes of data, too. Data mining offers the possibility of extracting previously unknown and potentially useful knowledge and patterns from large databases. The process consists of numerous steps such as integration of data from several databases or data warehouses, preprocessing of the data and induction of a model with a learning algorithm. The model is then used to identify and implement some decisions within the company.

Strong competition on the market is forcing organizations to identify innovative ways to increase their market share while reducing cost. Therefore, data mining can provide significant competitive advantage to a company by exploiting the potential of large data warehouses. It has an important role in helping the companies to understand their clients' behavior, anticipating the stocks, to optimize the sales policy and other benefits.

The aim of this technology is usually to find hidden but significant relationships that can lead to a bigger profit. The essential difference between the data mining techniques and traditional methods with databases is that in the second case, the database become passive and it is used only for storing large data amounts. In the first case, the database is no longer passive. Through an automated process of data analysis, it could offer useful information for the business plans.

The process of data mining involves multiple steps (see fig. 1). It starts with the selection of data incorporated in a training set that consists of observed values of certain attributes, generally historical data. The selected data are then cleaned and preprocessed. Cleaning is made in order to remove the discrepancies and preprocessing is responsible for consolidation of relevant information to the mining algorithm, trying to reduce the problem complexity. Among the steps in preprocessing, attribute selection has a special role. The data set is then analyzed to identify patterns, so that different inductive learning algorithms are applied. The model is finally validated with new data sets to ensure its generalizability. The steps in the mining process are performed iteratively until meaningful business knowledge is extracted.

An important issue here is the attribute selection which is preferable to be done before applying the learning algorithm. This involves a process for determining which attributes are relevant in that they predict or explain the data, and conversely which attributes are redundant or provide little information. A subset of M attributes out of N is chosen, complying with the constraint $M \leq N$, in such a way that characteristic space is reduced according to some criterion. Attribute selection guarantees that data getting to the mining phase are of good quality [1]. Identifying and keeping the attributes which are relevant to the decision making often provides valuable structural information.

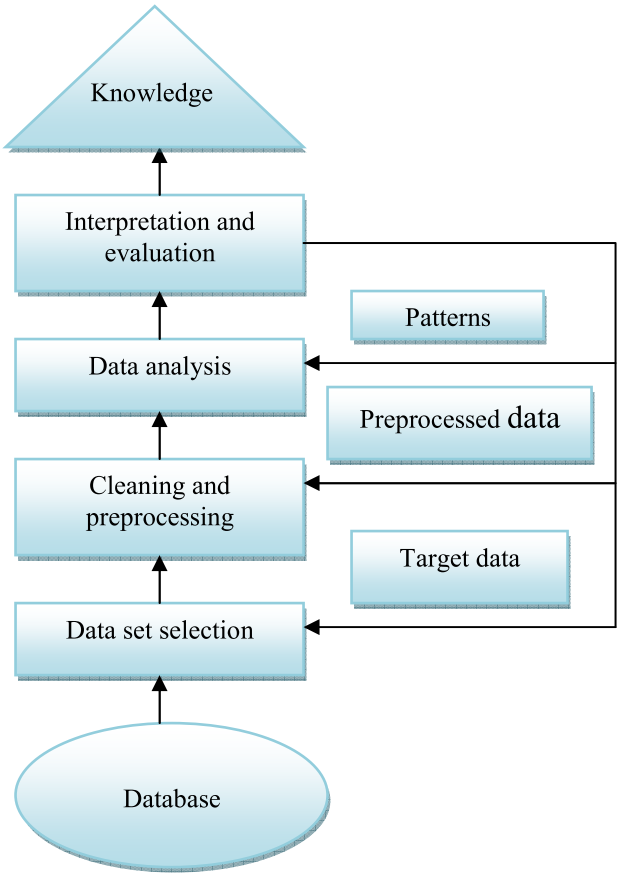

Fig. 1 - An overview of the data mining process
Source: adapted from U. Fayyad, G. Piatetsky, P. Smyth, 1996

Many attribute selection methods are based on optimization approach that includes genetic algorithms (Yang and Honavar, 1998), logical analysis of data (Boros, 2000), mathematical programming [2] (Bradley, 1998). One of the most efficient optimization methods for data mining is support vector machines or kernel methods and the most common concepts learned in data mining are classification, clustering and association. The following section includes a data mining application, namely customer relationship management systems (CRM).

II. Optimization Algorithms for Data Mining- Support Vector Machines (SVM)

SVMs were developed by Cortes & Vapnik (1995) for binary classification and they are based on the Structural Risk Minimization principle from computational learning theory. The SVM technique has been applied in many financial applications recently, mainly in the area of time series prediction and classification.

The algorithm is considering the following steps:

a) Class separation: we must find the optimal separating hyperplane between the two classes. Linear programming can be used to obtain both linear and non-linear discrimination models between separable data points or instances (Mangasarian, 1965). The problem is to determine a best model for separating the two classes.

If this hyperplane exists, then there are many such planes. The optimal separating hyperplane is the one that maximizes the sum of the distances from the plane to the closest positive example and the closest negative example.

b) Overlapping classes: data points on the wrong side of the discriminant margin are weighted down;

c) Nonlinearity: when we cannot find a linear separator, data points are projected into a higher-dimensional space where the data points effectively become linearly separable (this projection is realised via kernel techniques);

d) Problem solution: the whole task can be formulated as a quadratic optimization problem which can be solved by specific techniques.

A program able to perform all these tasks is called a Support Vector Machine.

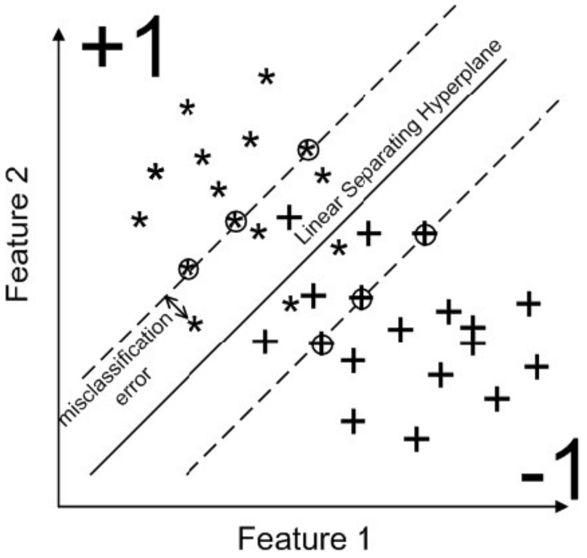

Fig. 2 - Illustration of a support vector machine in a two dimensional feature space.
Source: P. Watanachaturaporn, P. K. Varshney, Manoj K. Arora, Evaluation of factors affecting support vector machines for hyper spectral classification

Here is a briefly description of a SVM method:
There is an input space, denoted by X, an output space, denoted by Y, and a training set, denoted by S:

$$S=((x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)) \subseteq (X \times Y)^l$$

where l is the size of the training set.

SVM belongs to the type of maximal margin classifier, in which the classification problem can be represented as an optimization problem. The hyper plane H can be defined in terms of its unit normal w and its distance b from the origin. So, H = { x € Rm : x · w + b = 0}, where x · w is the dot product between two vectors. The aim of support vector machines is to orientate this hyper plane in such a way as to be as far as possible from the closest members of both classes.

$$\min_{w,b} < w, w >$$

$$s.t. y_i(< w, \phi(x_i) > +b) \geq 1,$$

$$i = 1, \ldots, l$$

Vapnik showed how training a support vector machine for pattern recognition leads to a quadratic optimization problem with bound constraints and one linear equality constraint:

$$\max W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < \phi(x_i), \phi(x_j) >$$

$$= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) s.t. \sum_{i=1}^{l} y_i \alpha_i$$

= 0, αi > 0, i = 1, … l.

where a kernel function, K(xi, xj), is applied to allow all necessary computations to be performed directly in the input space (a kernel function K(xi, xj) is a function of the inner product between xi and xj, thus it transforms the computation of inner product < /(xi), /(xj)> to that of < xi, xj >). Conceptually, the kernel functions map the original data into a higher-dimension space and make the input data set linearly separable in the transformed space. The choice of kernel functions is highly application-dependent and it is the most important factor in support vector machine applications (Z. Huanga et. al 2004).

Let's formulate the dual program:

$$\max W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < \phi(x_i),$$

$$\phi(x_j) >= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) s.t. \sum_{i=1}^{l} y_i \alpha_i$$

0 ≤ αi ≤ C, C is a constant that measures the penalty, i = 1, … l.

(Z. Huanga et. al 2004).

The standard SVM formulation solves only the binary classification problem, so we need to use several binary classifiers to construct a multi-class classifier or make fundamental changes to the original formulation to consider all classes at the same time.

III. Specific Operations and Techniques in Data Mining

As mentioned in the introduction, data mining is a technique that consists in analysing large volumes of information stored in data warehouses, in order to resolve decision problems. The technique is derived from three categories of software applications: the statistical ones, artificial intelligence applications based on neuro-fuzzy algorithms and the ones based on automated machine learning. Once the data has been prepared, the next step is to generate previously unknown patterns from the data using inductive learning. The most popular types of patterns are classification models, clustering and association rules that describe relations between attributions. These operations are detailed in the next sections.

III.1. Classification

Classification means grouping the observations/ cases based on a predictable attribute. Each case contains a set of attributes, of which one is the classification attribute (the predictable attribute). The operation consists in finding a model that describes the predictable attribute as a function of other attributes taken as input values. Basically, the objective of the classification is to first analyze the training data and develop a model for each class using the attributes available in the data. Such class descriptions are then used to classify future independent test data or to develop a better description for each class.

In order to train a classification model, the class values of each case in the data set must be known, values that are usually found in historical data.

The most common classification algorithms include support vector machines (see Section 2), decision trees, neural networks and Bayesian networks.

III.1.1. Decision trees

The basic idea of this algorithm is to split recursively a set in subsets containing more or less homogeneous states of the variables, i.e. those whose prediction we are interested in. The first iteration process the root node which contains all the data. The next iterations are working on derived nodes which contain subsets. At each iteration, there is necessary to choose the independent variable which divides most effectively the data, so that the obtained subsets be as homogeneous in relation to the dependent variable.

A decision tree is a tree-like structure of decision diagrams, in which each node represents a test of an attribute, each branch is a test result, and the leaves are classes or class distribution.

There are several algorithms based on decision trees:

- CHAID (Chi-square automatic interaction detection);
- CART (Classification and regression tree);
- C4.5, where the attributes are chosen to maximize the information gain ratio in the split (Quinlan, 1993).

Algorithms such CART and C4.5 are computationally efficient and proved to be very successful in practice.

CART is an algorithm for exploration and prediction and it chooses each predictor when building the tree in order to reduce data clutter. The measure on which is preferred one predictor to another is the entropy value. CART algorithm is relatively robust against missing data. If a value is missing for a particular predictor in a particular record, in the construction process of the tree, that record will not be used in making the determination of optimal branching. When

CART is used to predict the new data, missing values can be manipulated by its surrogates. Surrogates are ramification values and predictors which simulate real tree branch and can be used when the data for the desired predictors is missing.

C4.5 works in three main steps. First, the root node at the top node of the tree considers all samples and passes through the samples information in the second node called 'branch node'. The branch node generates rules for a group of samples based on entropy measure. In this stage, C4.5 constructs a very big tree by considering all attribute values and finalizes the decision rule by pruning. It uses a heuristic approach for pruning based on statistical significance of splits. After fixing the best rule, the branch nodes send the final target value in the last node called the 'leaf node' (S. Ali, K.A. Smith, 2006).

In recent studies, Street presents a new algorithm for multi-category decision tree induction based on non-linear programming called OC-SEP (Oblique Category SEParation).

The generation of a DT involves partitioning the model data set into at least two parts: the training data set and the validation data set (commonly referred to as the test data set). There are two major phases of the DT generation process: the growth phase and the pruning phase (Kim H, Koehler, 1995).

The growth phase involves inducting a DT from the training data such that either each leaf node is associated with a single class or further partitioning of the given leaf would result in the number of cases in one or both child nodes being below some specified threshold. The pruning phase aims to generalize the DT that was generated in the growth phase in order to avoid over fitting. Therefore in this phase, the DT is evaluated against the test (or validation) data set in order to generate a subtree of the DT generated in the growth phase that has the lowest error rate against the validation data set. It follows that this DT is not independent of the training data set or the validation data set (i.e. commonly called test data set). For this reason it is important that the distribution of cases in the validation data set correspond to the overall distribution of the cases (Kweku-Muata, Osei-Bryson, 2004).

III.1.2. Neural networks

A neural network consists of at least three layers of nodes. The input layer consists of one node for each of the independent attributes. The output layer consists of node(s) for the class attribute(s), and connecting these layers is one or more intermediate layers of nodes that transform the input into an output. When connected, these layers of nodes make up the network we refer to as a neural net. The training of the neural network involves determining the parameters for this network. Specifically, each arc connecting the nodes in this network has certain associated weight and the values of those weights determine how the input is transformed into an output. Most neural network training methods, including back-propagation, are inherently an optimization processes. As before, the training data consists of values for some input attributes (input layer) along with the class attribute (output layer), which is usually referred to as the target value of the network (S. Olafsson et al., 2008) The optimization process seeks to determine the arc weights in order to reduce some measure of error (normally, minimizing squared error) between the actual and target outputs (Ripley, 1996).

The most popular neural net algorithm is back-propagation. It performs reverse processing, starting from the output node and calculating each node‘s contribution to all previous error. Thus, not only that it is possible to calculate the contribution of each node, but of the weights to each error. In this way, the error is propagated back through the entire network, resulting in adjustments to the weights that contributed to the error.

This cycle is repeated for each case from the learning set, being made slight change of the weights after each processed case. When the entire set of learning has been processed, it will be processed again. Each scroll of the entire set is called epoch. It is possible that the network training to require hundreds or even thousands of epochs. This way, even if the processing of a case lead to minor modifications of the weights, in the end the changes are significant because of the accumulation effect.

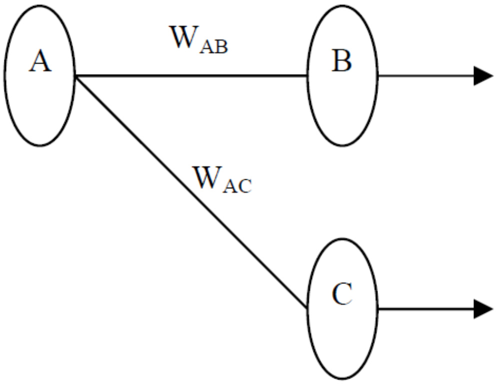

Fig 3 - A single connection learning in a Back Propagation network

Neural networks algorithms use different stopping rules to control the end of the network training process**.** Examples of rules commonly encountered are:

- Stop after a specified number of epochs;
- Stop when the error falls below a predetermined value;
- Stop when the error was not reduced after a certain number of epochs.

III.1.3. BAYESIAN NETWORKS

Bayesian networks involve conditional independence of the classes, and in practice, it shows the highest accuracy of all methods of classification. But in practice there are often dependencies between variables. Bayesian networks show in a graphical manner the dependencies between variables. Based on the network structure, it can realize various kinds of inferences. In a Bayesian network:

- nodes represent random variables
- oriented ties: X has a direct influence on Y each node has an associated conditional
- probability table that quantifies the effect of parents on the node.

There are two key optimization-related issues when using Bayesian networks. First, when some of the nodes in the network are not observable, that is, there is no data for the values of the attributes corresponding to those nodes, finding the most likely values of the conditional probabilities can be formulated as a non-linear mathematical program. In practice this is usually solved using a simple steepest descent approach. The second optimization problem occurs when the structure of the network is unknown, which can be formulated as a combinatorial optimization problem (S. Olafsson, 2008).

The inference algorithm is as follows:

To use a Bayesian network as a classifier, one simply calculates argmaxy P(y|x) using the distribution P(U) represented by the Bayesian network. Thus:

$$\begin{aligned} P(y|\mathbf{x}) &= P(U)/P(\mathbf{x}) \\ &\propto P(U) \\ &= \prod_{u \in U} p(u|pa(u)) \end{aligned}$$

Since all variables in x are known (x = x1….xk, are attribute variables), we do not need complicated inference algorithms, but just calculate the formula for all class values (R. Bouckaert, 2004).

The absence of arcs in a Byesian networks involves conditional independence relations which can be exploited to obtain efficient algorithms for computing marginal and conditional probabilities. For single connected networks, in which the underlying undirected graph has no loops, there is a general algorithm called belief propagation. For multiply connected networks, in which there can be more than one undirected path between any two nodes, there exists an algorithm known as junction tree algorithm.

III.2. CLUSTERING

Webster (Merriam-Webster Online Dictionary, 2008) defines cluster analysis as ‘‘a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics.

The scope of the cluster analysis is to search and identify the classes, groups or clusters within sets of objects or forms, so that the elements belong to the same class to be as similar and elements belonging to different classes are as different among them. In other words, cluster analysis is a modality to examine the similarities and non- similarities of objects belonging to a certain set. The goal is to group these objects in order to form distinct and internally homogeneous classes. Clustering is different from classification because it does not require predefined classes. The records are grouped based on self-similarity. The algorithms can be divided into two categories: hierarchical clustering and partitioned clustering.

Partitional algorithms process the input data and create a partition that groups the data into clusters. In contrast, hierarch- ical algorithms build a nested partition set called a cluster hierarchy. At the lower level of the hierarchy, all the data points are usually clustered in singleton clusters, i.e.clusters composed of just one object or data point.The top most partition is usually the cluster containing all the data points.The remaining intermediate partitions define the path followed from one end to the other. The procedures to obtain this set of partitions are classified as either agglomerative (bottom-up) or divisive (top-down). The former begins from the singleton clusters and obtains the hierarchy by successively merging clusters. In contrast, the latter begins with a single cluster containing all the points and proceeds by iteratively splitting the clusters (I. Gurrutxaga et al, 2010).

Partitioned algorithms run much faster than hierarchical ones, which allow them to be used in analyzing large datasets, but they have their disadvantages as well. Generally, the initial choice of clusters is arbitrary, and does not necessarily comprise all of the actual groups that exist within a dataset. Therefore, if a particular group is missed in the initial clustering decision, the members of that group will be placed within the clusters that are closest to them, according to the predetermined parameters of the algorithm. In addition, partitioned algorithms can yield inconsistent results- the clusters determined this time by the algorithm probably won't be the same as the clusters generated the next time it is used on the same dataset.

In the early literature, Vinod (1969) writes two integer programming formulations of the clustering issue. In the first formulation, the decision variable is defined as an indicator of the cluster to which each instance is assigned and the goal is to minimize the total cost of the assignment:

$$x_{ij} = \begin{cases} 1 & \text{if the } i\text{th instance is assigned} \\ & \text{to the } j\text{th cluster,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{k} x_{ij} = 1, \quad i = 1, 2, \ldots, n \\ & \sum_{i=1}^{n} x_{ij} \geqslant 1, \quad j = 1, 2, \ldots, k. \end{aligned}$$

Where wij is some cost of assigning the ith instance to the jth cluster.

In this kind of optimization formulations , clustering is defined by fixed number of centres.

Bradley and Mangasarian (2000), use a different definition of a clustering and instead of finding the best centres, they identify the best cluster planes:

$$P_j = \{x \in \mathbf{R}^m | \mathbf{x}^{\mathrm{T}} \cdot \mathbf{w}_j = \gamma_j\}, \quad j = 1, 2, \ldots, k.$$

The authors propose an iterative algorithm that assigns instances to the closest cluster and then given the new assignment finds the plane that minimizes the sum of squares of distances of each instance to the cluster. In other words, given the set of training instances A(j) € Rn assigned to cluster j, find w and γ that solve:

$$\begin{aligned} \min_{\mathbf{w},\gamma} \quad & \|\mathbf{A}\mathbf{w} - \mathbf{e}\gamma\|_2^2 \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}} \cdot \mathbf{w} = 1. \end{aligned}$$

(w*j, γ j*) is the optimal solution by letting w*j be the eigenvector corresponding to the smallest eigenvector of (A (j))T(I – e · eT/nj) A(j), where nj = |A(j) | is the number of instances assigned to the cluster, and: γ j* = eT A (j) w*j / nj.

In addition to the above models, Kroese et al. (2004) used the cross-entropy method to solve both discrete and continuous versions of the problem. They show that although the cross-entropy method is more time consuming than traditional heuristics such as k-means the quality of the results is significantly better.

Data clustering has been used for the following three main purposes:

- Underlying structure: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features.
- Natural classification: to identify the degree of similarity among forms or organisms (phylogenetic relationship).
- Compression: as a method for organizing the data and summarizing it through cluster prototypes.

III.3. ASSOCIATION RULES

The goal of association rule mining is to find all rules satisfying some basic requirement, such as the minimum support and the minimum confidence. It was initially proposed to solve market basket problem in transaction databases, and has then been extended to solve many other problems such as classification.

The process is described as follows:

Let I = {i1, i2, . . ., it}, refers to the set of literals called set of items and the set D = {t1, t2, . . ., tk}, refers to the transactional dataset. Let antecedent A and consequent B be item sets, where $A, B \subseteq I$. An association rule is an implication of the form $A \Rightarrow B$., where A ∩ B = Ø. The support of an association rule $A \Rightarrow B$., denoted sup ($A \Rightarrow B$.), is defined as a number of transactions in D containing both A and B. The confidence of an association rule $A \Rightarrow B$.,, denoted cfi ($A \Rightarrow B$.) is the conditional probability that an instance contains item set B given that it contains item set A and is defined as :

conf ($A \Rightarrow B$.) = sup (A $\cup$ B)/ sup (A).

Support and confidence are typically modelled as constraints for association rule mining, where users specify the minimum support supmin and minimum confidence confmin according to their preferences. An item set is called a frequent item set if its support is greater than this minimum support threshold.

Apriori is the best-known and original algorithm for association rule discovery (Agrawal and Srikant, 1994). "The idea behind the apriori algorithm is that if an item set is not frequent in the database then any superset of this item set is not frequent in the same database. There are two phases to the inductive learning: (a) first find all frequent item sets, and (b) then generate high confidence rules from those sets. The apriori algorithm generates all frequent item sets by making multiple passes over the data. In the first pass it determines whether 1-item sets are frequent or not according to their support. In each subsequent pass it starts with those item sets found to be frequent in the previous pass. It uses these frequent items sets as seed sets to generate super item sets, called candidate item sets, by only adding one more item. If the super item set meets the minimum support then it is actually frequent. After frequent item sets' generation, for each final frequent item set it checks all single-consequent rules. Only those single consequent rules that meet minimum confidence level will go further to build up two-consequent rules as candidate rules. If those two-consequent rules meet the minimum confidence level will continue to build up three-consequent rules, and so on" (S. Olafsson et al, 2008).

Despite the great achievement in improving the efficiency of mining algorithms, the existing association rule models used in all of these studies incur some problems. First, in many applications, there are taxonomies (hierarchies), explicitly or implicitly, over the items. It may be more useful to find association at different levels of the taxonomy than only at the primitive concept level (J. Han, R, Srikant).

Second, the frequencies of items are not uniform. Some items occur very frequently in the transactions while others rarely appear. A uniform minimum support assumption would hinder the discovery of some deviations or exceptions that are more interesting but much less supported than general trends. Furthermore, a single minimum support also ignores the fact that support requirement varies at different levels when mining association rules in the presence of taxonomy (Tseng, 2007).

Rastogi and Shim (2001) consider that optimized association rules are an effective way to focus on the most interesting characteristics involving certain attributes. Optimized association rules are permitted to contain un-instantiated attributes and the problem is to determine instantiations such that either the support, confidence or gain of the rule is maximized.

They are useful for unravelling ranges for numeric attributes where certain trends or correlations are strong (that is, has high support, confidence or gain). The authors used dynamic programming to generate the optimized support rules.

The optimization problem maximizes the gain subject to minimum support and confidence:

Max gain ($A \Rightarrow B$)
s.t. sup($A \Rightarrow B$) ≥ supmin
conf ($A \Rightarrow B$) ≥ confmin

where gain ($A \Rightarrow B$) = sup {(A1 € [v1, v2]) □ C1} - confmin · sup(C1) = sup ($A \Rightarrow B$) · (conf($A \Rightarrow B$ - confmin)

where A is a numeric attribute, v1 and v2 are a range of attribute A and C1 is a normal attribute.

IV. CUSTOMER RELATIONSHIP MANAGEMENT SYSTEMS OPTIMIZATION BY USING DATA MINING TECHNIQUES

A customer relationship management system (CRM) is a bucket of IT applications and procedures whose target is to identify the main expectations and preferences of the clients and to use efficiently the gathered information in order to improve the relationships between the business and the customers. The implementation of such system implies two components:

- The managerial component, consisted of the total methods and techniques used for the integration and usage of data related to the customers behaviour;
- The IT component, which includes the hardware and software equipment used for data collection, storing and management.

The main components of CRM systems are:

- A stop shop is the input point in the system for the data, meaning the requests and claims of the clients, which then are processed within a documents management system process;
- Contact Center/ Help Desk offers special assistance to the clients who ask for information regarding the specific products and services. Developing such a component provides many advantages: reducing the number of missed calls by intelligent distribution of calls, increasing the productivity of the marketing and sales departments, enhancing customer satisfaction by increasing the value that he perceives, monitoring the satisfaction of customers.
- eCRM meaning the internet technology using specific instruments, such as: personalized e-mail addresses, chat or interactive dialogs, forums.

According to [3], CRM consists of four dimensions:
(1) Customer Identification;
(2) Customer Attraction;
(3) Customer Retention;
(4) Customer Development.

They share the common goal of creating a deeper understanding of customers to maximize customer value to the organization in the long term. Data mining techniques, therefore, can help to accomplish such a goal by extracting or detecting hidden customer characteristics and behaviours from large databases.

The main advantages of CRM implementation are: more efficient activities of the orders received from consumers, improving the quality of services provided to the clients, a qualitatively higher level communication with the client by using multiple communication channels (telephone, stop shop, web, e-mail), reducing the communication costs with clients, reduce time consuming for claims, achieving a better image of the organization in front of clients.

In practice, especially in the large companies, applying CRM techniques implies the following steps:

1. Identify the organization's clients and including them in different categories depending on their preferences and behaviors. We can split the clients in four categories:

a) Clients with general requirements and an uniform character;
b) Clients with specific requirements and an uniform character;
c) Clients with general requirements and no uniform character;
d) Clients with specific requirements and no uniform character

2. Establishing the necessary information and design the system architecture. In this phase, there is planning the clients management database which includes, in general, information related to: identification of person, professional training, social status, embership in a particular category of clients, attitudes and perceptions, behaviors in different situations, requests, complaints submitted by customer.

3. Identifying ways of information gathering which involves developing a toolbox of methods and techniques whereby information describing customer behaviors to be collected and entered into the database.

4. Gathering information and updating the database that consists of applying the techniques defined in the second stage, with the scope of the consolidation of customer database.

5. Operationalisation of changes in the organizational plan for enhancing the customer satisfaction by improving and diversifying provided services, acting simultaneously both in terms of coverage general requirements and individual ones. Studies reveal that the amplification of satisfaction degree generates an improved image of the organization on the market, but only up to a maximum point, beyond which the image begins to deteriorate.

Data mining plays an important role in CRM by identifying customer behaviour patterns from customer usage data and predicting which customers are likely to respond to cross-sell and up-sell campaigns, which is very important to the business [4]. Regarding former customers, data mining can be used to analyze the reasons for churns and to predict churn [5].

Optimization also plays an important role in CRM and in particular in determining how to develop proactive customer interaction strategy to maximize customer lifetime value. A customer is profitable if the revenue from this customer exceeds company's cost to attract, sell and service this customer. This excess is called the customer lifetime value [6].

E.W.T. Ngai in [7] proposes a graphical classification framework on data mining techniques in CRM as shown in figure 2:

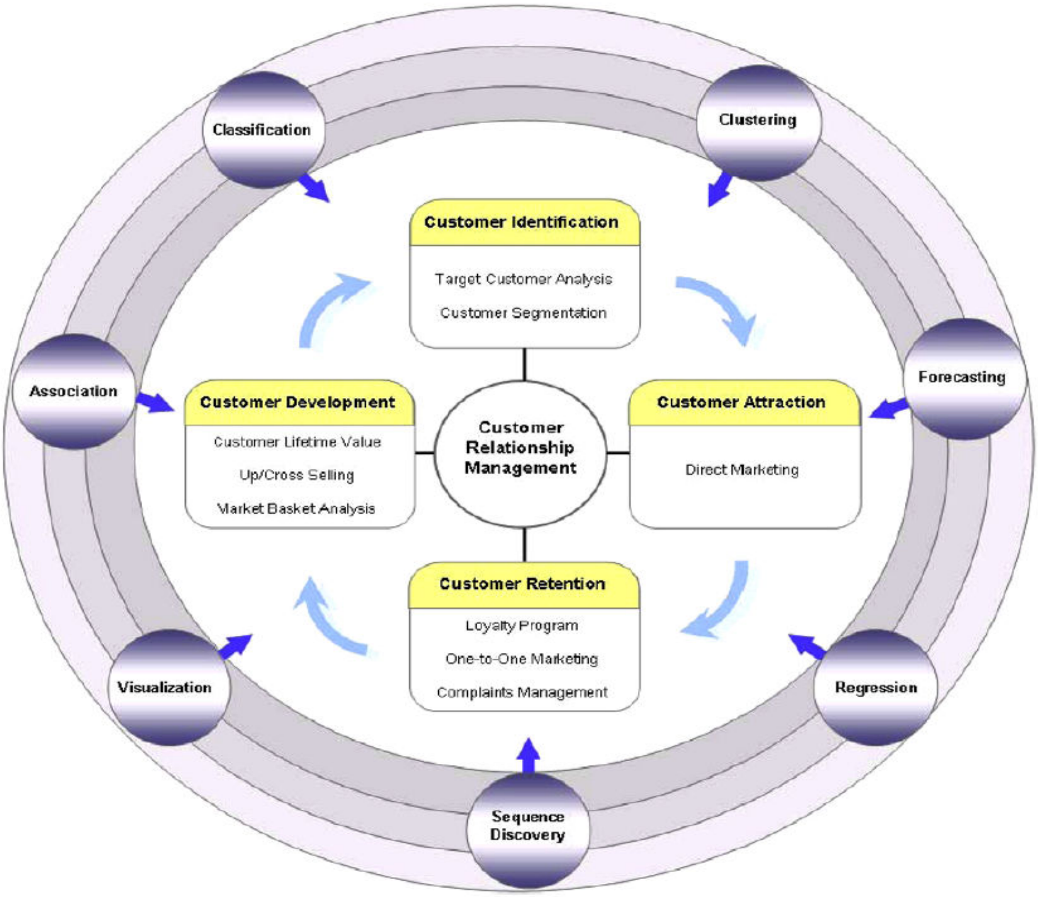

Fig. 2 - Classification framework for data mining techniques in CRM.

Source: [7] E.W.T. Ngai, Li Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, 2009.

Data mining techniques can be used successfully, especially because CRM implies a multidimensional approach which can, by instance, include three dimensions:

- hierarchy of products (brand, class, category, product);
- hierarchy of periods (years, quarters, months, dates);
- customer hierarchy (regions, areas class customers).

In practice, this approach is successfully performed through a modern concept that stands today in the majority of process support systems decision, namely the OLAP (On-Line Analytic Processing), which is based on technical
multidimensional data analysis [9], [10].

If we refer to the CRM user demands, OLAP systems provide support for real-time satisfaction of specific claims, because they anticipate the timing and content of the interrogation and provides the optimal combination between pre-calculated results and those calculated at the time of information requested. OLAP systems use a specific tool, reason for which most experts believe that they represent the best environment for implementation of functional information models based on systems dynamic principles.

At present, almost all large organizations hold an intranet platform which, together with some extensions and instruments, provide the basic functionality of Business Intelligence applications, such as organizing information in data warehouses and processing them using data mining techniques. Numerous specific data mining functions are already implemented as components of the Intranet architecture or like specific solutions such as CRM.

V. Conclusions

The large volume of information that decision makers are facing, requires advanced processing technologies, but also new types of systems to assist decision. Business Intelligence is currently offering solution for the problems in decision making at all managerial levels.

Data Mining, as part of BI systems, has enjoyed great popularity in recent years, with advances in both research and commercialization. Data mining is focused on assessing the predictive power of models and performs analysis that would be too hard-working and time-consuming by using traditional statistical methods. It offers important information which is used to improve customer retention, response rates, attraction, and cross selling. As shown in the paper, through the full implementation of a CRM program, the companies increase the value of their customers, keeping and attracting the right ones.

Although many books and articles have been written on Business Intelligence topic, it still represents a promising research field. Interest in data mining continues to increase and the potential of using optimization methods needs more study. Also, investigating how to combine optimization and data mining techniques, especially in the CRM area, should be encouraged for many reasons. Data mining and optimization can be integrated to build customer profiles, which is absolutely necessary in many CRM applications.

Acknowledgements

This article is a result of the project POSDRU/88/1.5./S/55287 „Doctoral Programme in Economics at European Knowledge Standards (DOESEC)" . This project is co-funded by the European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies in partnership with West University of Timisoara.

Also, this paper presents some preliminary researches of the project PN II, TE Program, Code 332: “Informatics Solutions for decision making support in the uncertain and unpredictable environments in order to integrate them within a grid network”, financed within the framework of People research program.

References

[1] Liu, H., Motoda, H., Feature Selection for Knowledge Discovery and Data Mining, Kluwer academic Publishers, 1998.

[2] Bradley, P.S., Mangasarian, O.L., k-Plane clustering. Journal of Global Optimization 16 (1), 23–32, 2000.

[3] Kracklauer, A. H., Mills, D. Q., & Seifert, D. Customer management as the origin of collaborative customer relationship management. Collaborative Customer Relationship Management - taking CRM to the next level, 3–6, 2004.

[4] Chiang, I., Lin, T., Using rough sets to build-up web-based one to one customer services. IEEE Transactions, 2000.

[5] Chiang, D., Lin, C., Lee, S., Customer relationship management for network banking churn analysis. In: Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, NV, 135–141, 2003.

[6] Sigurdur Olafsson, Xiaonan Li, Shuning Wu, Operations research and data mining, European Journal of Operational Research 187, 1429–1448, 2008.

[7] E.W.T. Ngai, Li Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications 36, 2592–2602, 2009.

[8] Parvatiyar, A., & Sheth, J. N. Customer relationship management: Emerging practice, process, and discipline. Journal of Economic & Social Research, 3, 1–34, 2001.

[9] Bâra A., Lungu I., Oprea S. V. - *Public Institutions' Investments with Data Mining Techniques,* Journal WSEAS Transactions on Computers, Volume 8, 2009, ISSN: 1109-2750, http://www.worldses.org/journals/computers/computers-2009.htm

[10] Bâra A., Lungu I., Velicanu M., Oprea S.V. - *Intelligent Systems for Predicting and Analyzing Data in Power Grid Companies*, TheProceedings of the IEEE International Conf. on Information Society (i-Society 2010) London, july 2010.

[11] Muntean M, Bologa AR, Bologa R, Florea A - Business Intelligence Systems in Support of University Strategy, Proceedings of the 7th WSEAS/IASME Int. Conf. on Educational Technologies, p. 118-123, WSEAS Press, 2011, ISBN 978-1-61804-010-7

[12] Khlif W, Zaaboub N, Ben-Abdallah H - Coupling Metrics for Business Process Modeling, WSEAS TRANSACTIONS on COMPUTERS, Volume 9, 2010, ISSN: 1109-2750

[13] Yang J, Hongjian Qu , Zhou L - Research on the Evaluation Methods of Bid of Construction Project Based on Improved BP Neural Network, WSEAS TRANSACTIONS on COMPUTERS, Volume 9, 2010, ISSN: 1109-2750