UNIVERSITY OF NIŠ
FACULTY OF ECONOMICS
"**ECONOMIC THEMES**"
Year XLVI, N° 4, 2008, p. 101-115
Address: Trg kralja Aleksandra Ujedinitelja 11, 18000 Niš
Phone: +381 18 528 601 Fax: +381 18 523 268

USING DATA MINING TECHNIQUES IN MARKET RESEARCH

Vinko Lepojević[*]
Vesna Janković-Milić[*]

Abstract: *Data mining techniques have recently gained popularity with researchers, in part because they overcome many limitations of traditional statistics and can handle complex data sets. Data mining is a process of extracting previously unknown, valid, actionable, and ultimately comprehensible information from large databases and then using the information to make crucial business decisions. Data mining techniques offer a powerful complement to statistical techniques and have useful research applications for customer satisfaction.*

Keywords: *Data mining, market research, data visualization, association rules, case-based reasoning, neural networks, genetic algorithms*

1. Introduction

Like any other discipline, marketing is concerned with finding predictable relationships among variables. What are the characteristics of loyal consumers? What attitudes and behaviours lead to a purchase? Why do people switch brands? What attributes and characteristics of a customer contribute to his or her creditworthiness?

Human behaviour being complex, such questions are not easily answered. For instance, customer loyalty may involve a number of factors such as a person's age, gender, place of residence, income level, marital status, availability of alternatives, and past purchase patterns, to name a few. Even when we know the factors that influence purchase behaviour, there is

[*] Faculty of Economics Niš; e-mail: vinko@eknfak.ni.ac.rs ; vesnajm@eknfak.ni.ac.yu
UDC 004.6:339.13.017
Received: September 12, 2008

still the problem of knowing the combination of characteristics that would best predict what we are interested in – in this example, customer loyalty. The major problem here is that thousands or even millions of combinations of predictor varibles are possible. It is impossible to sift through all possible combinations of relationships except by mechanical means.

Data mining is the mechanical search for patterns and relationships in data. Consider these examples:

1. A large bank has considerable information on its customers. A customer asks for a large loan. From the characteristics of the customer – age, income, place of residence, marital status, number of children, net worth, date of inception of account, average monthly balance, etc. – can the bank calculate beforehand how risky it is to lend money to this customer?

2. A supermarket chain wants to assess what non-baby products it should stock in the baby products aisle, so it can increase the sale of unrelated items when a customer comes to buy baby products. This means understanding the patterns of purchase by thousands of customers of thousands of products.

3. A hotel chain would like to understand the patterns of guest registration and behaviour so it can offer discounted rates at slow period to attract additional sales without greatly reducing its revenue through discounting.

4. A telephone company is interested in managing its customer relationships based on individual customer characteristics.

From a broad perspective, data mining involves all of the following:

- data collection (data warehousing, web crawling);
- data cleaning (dealing with outliers, errors);
- feature extraction (identifying attributes of interest);
- pattern discovery and pattern extraction;
- data visualization;
- results evaluation.

However, from a more focused perspective, the term “data mining” is generally applied to pattern discovery and pattern exstraction.

Conceptually, data mining applications in marketing fall into two broad categories – grouping of variables and identifying functional relationships among variables. While the basic multivariate techniques discussed in this book thus far cover these two aspects, data mining techniques are concerned with achieving these objectives with potentially large databases with a large number of variables. Condensing a large number

of variables into potentially meaningful groups and finding relationships among a large number of variables- often without any underlying hypotheses – characterize these techniques. In that they resemble exploratory data analysis.

Although data mining may be the only available way to deal with the analysis of complex patterns, it is best to consider it as an exploratory technique to generate hypotheses rather than as a confirmatory technique that leads to conclusions. As Bonferroni's theorem warns us, if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no, physical validity.

2. Data mining models

Data mining uses a variety of analytic tools to uncover patterns in data. Although there are many such tools, the following are the ones most frequently used:

- Data visualization
- Association rules
- Case-based reasoning
- Neural networks
- Genetic algorithms

2.1. Visualization

Data visualization takes advantage of the capacity of human beings to recognize and distinguish patterns of observable characteristics. Visualization is particularly effective for exploring and condensing large amounts of messy data into compact understandable pictures. These techniques range from exploratory techniques such as simple histograms, box plots, scatter diagrams, and link analysis networks to more complex techniques such as rotating multicolored three-dimensional surface plots in three dimensions.

2.2. Association rules

Association rules state the relationships between the attributes of a group of individuals and one or more aspects of their behaviour. The purpose of these rules is to enable predictions about the behaviour of other individuals who are not in the group but possess the same attributes. Association rules are stated in dichotomous terms such as good credit risk vs. bad credit risk, buyer vs. non-buyer. These rules assign probability – like numbers to actions.

Association rules are of the form $\{X_1, X_2, ..., X_n\} \Rightarrow Y$: if we find all of $X_1, X_2, \ldots, X_n$, then we have a high probability of finding Y. As an example of an association rule, suppose a mail-order institution is interested in cross-selling a personal digital assistant to those who have just ordered several electronic items. Promoting a PDA to unlikely customers may antagonize them while wasting the salespeople's time. Therefore, the company would like to restrict the offer to customers who have a high probability of buying a PDA. To accomplish this the company can analyze purchase who bought a CD player and a wireless telephone on one call where much more likely to buy a PDA on a subsequent call than customers who ordered tape recorders or calculators. Consequently, when the association rules are incorporated in the company's order entry system and the system identifies that the customer on the phone recently ordered a CD player and a wireless telephone, it prompts the sales person to make the offer on a PDA. On the other hand, if the system finds that the caller bought a tape recorder or a calculator on the last order, the built-in decision rules will not prompt the salesperson to make that offer (but presumably an offer of another product which has a high probability of purchase for those who bought tape recorders or calculators).

The probability level of finding Y for us to accept this rule is known as the confidence of the rule. Generally, we would search only for rules whose confidence is above a certain threshold and is significantly higher than what would be obtained if X s is chosen at random. The purpose of the latter condition is to avoid spurious associations. For instance, a supermarket might find a rule like $\{chocolate, cigarette\} \Rightarrow$ newspaper, but that might only be because a lot of people buy newspapers, irrespective of what else they might buy.

2.3. Three-based methods (decision trees)

Three-based methods are used to sequentially partition the data set using independent variables in order to identify subgroups that contribute most to the dependent variable. The most commonly used techniques for automatic sequential splitting are chi-squared automatic interaction detectors (CHAID) and classification and regression trees (CART). Three-based methods are good at identifying the most important variables, interactions among independent variables, and non-linear relationships. They help to identify the most important variables and eliminate the irrelevant ones. The results obtained using these methods are relatively easy for users to understand and interpret. Decision-tree algorithms are robust to outliers and erroneous data.

Consider a simple example of identifying customers of a bank who are likely to respond to a direct mail campaign; data are available on net worth, income and gender. (In practice, three-based methods in data mining are unlikely to be used when there are only three independent variables. However, the principle of splitting is the same, whether one uses 3 or 300 independent variables). Records show that 20% of those who were sent the direct mail responded. The objective of the analysis is to identify subgroups that will have a much higher probability of responding to the offer, based on the three independent variables. To keep things simple, let us group each of the independent variables into just three categories: net worth (high, low), income (high, low) and gender (male, female) (male, female). At the first lever of analysis, our question is which of these three independent variables differentiates responders from non-responders from non-responders. Suppose the data show that the response rates are as follows: among males 15%, and females 25%; among those with high net worth 35%, and low net worth 7%; among those with high income 30%, and low income 14%. Obviously, among the three variables, a person's net worth maximally differentiates between responders and non-responders the best. So customers are split into two groups; high net worth and low net worth. This process is repeated separately for the high net worth ant the low net worth. This process is repeated separately for the high net worth and the low net worth groups with the remaining two variables. When there are a number of independent variables, similar analysis is automatically performed at each stage process until a prespecified criterion is met (e.g. statistical significance or a minimum specified difference between the groups).

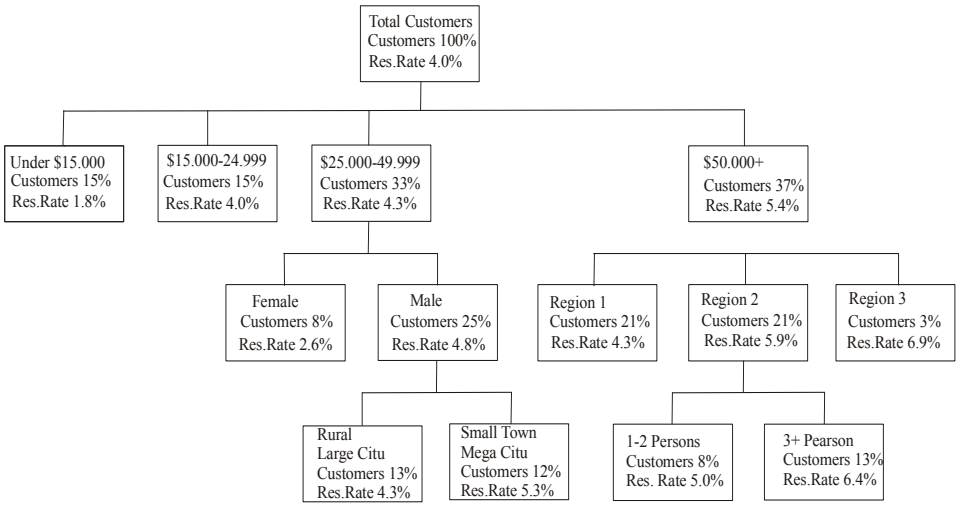

Exhibit 1 CHAID analysis

Exhibit 1 shows an example of this type of analysis. Each group can be profiled by following the three hierarchies. Decision trees provide a defined way to develop segments in terms of a single dependent variable. However, since decision trees split the sample sequentially, they use up data rapidly and, therefore, are not suitable for use in small databases. These techniques are highly sensitive to noise in the data and they tend to *overfit* data. As a result it is important to cross-validate the findings obtained using decision-tree results.

2.4. Case-based reasoning

In case-based reasoning (CBR) systems, we compare the attributes of a new case with corresponding attributes in a collection of previously known cases. The objective re is to identify examples that provide generally positive solutions and use them to generate a template for the current case. As an example, consider a fast food chain looking into setting up new outlets. Among the specifications of such outlets, it needs to consider such matters as floor space, number of counters and whether or not it include a salad bar. It will also need to compare the attributes of the locality in which any potential new outlet is situated, such as average income, the number of teens and preteens, traffic flow, and number of commercial establishments to the corresponding attributes for all of the company's existing outlets, along with their design specifications, annual sales and profitability. The CBR system is used to identify existing locations whose attributes most closely resemble those of the proposed locations and develop design specifications for the proposed outlets. Obviously this is not meant to be a mechanical exercise since any successful site may have some features that are unique. In other words, the characteristics of successful outlets are used as a template to be modified as required and not as a mould that is inflexible.

The value of case-based reasoning systems rests on the fact that it forces the user to focus on the similarities and differences between different situations in a structural way using the attributes that define the cases. In doing so CBR are easy to understand and implement on the computer. They accommodate qualitative and quantitative variables and can deal with discontinuous variables.

On the negative side, CBR's represent *what was actually done* in the past, not necessarily *what is optimal* under similar circumstances. The solutions of the past may not necessarily be optimal under current conditions, and using them to solve current problems may simply perpetuate mistakes and sub optimal solutions of the past. Establishing and maintaining CBRs will require considerable expertise and time investment. It is not a simple task to

first identify attributes that are related to specific outcomes and then to assign weights so that new situations can be matched to the most appropriate outcomes. Even more importantly, CBRs may lead to misleading conclusions when there are significant interactions among variables under consideration.

2.5. Neural networks

Neural networks are computer models that are designed to simulate human brain processes and are capable of learning from examples to find patterns in data. Although they have been around for decades, only recently have they begun to make an impact in marketing. This may be attributed to the rapidly reducing costs of computing and the emergence of better theoretical frameworks. Unlike in conventional computing, neural nets do not rely on specified methodology based on a standard set of instructions. Rather, a neural net is "trained by example", just as a rat learns a maze or a child learns to walk.

Supervised neural network

Conceptually, a neural net is a "black box" that produces a set of outputs on the basis of a set of inputs (Exhibit 2). The network is presented with a "training set" in which both the inputs and outputs are known. Using this as an example, the network is "trained" to model the outputs from the inputs. This is known as "supervised training".

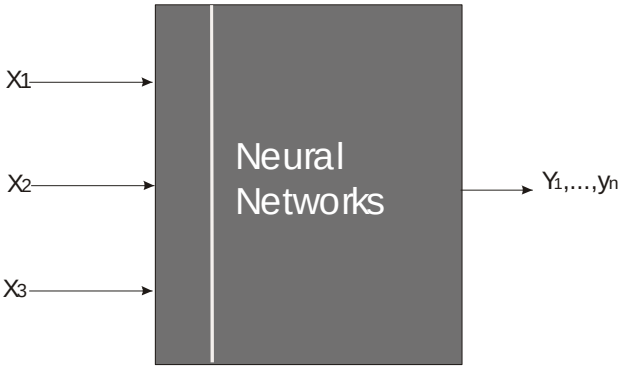

Exhibit 2 The neural network black box

The "black box" consists of a collection of processing units (analogous to neurons of the brain) that are connected and form a "network". "Training" involves varying the weights assigned to the connections between

neurons with a view to minimizing the difference between the network's outputs and the actual outputs obtained from the training set. The model thus generated is then validated by applying it to a separate "test set" of data.

The multiplayer perceptron

A common model of neural nets used in business is one known as the multiplayer perceptron (MLP). In the example shown in Exhibit 3, there are two inputs and one output, and there are three layers of neurons. The neurons in one layer are connected to every neuron in the next layer. The neuron takes the sum of its inputs and applies to every neuron in the next layer. The neuron takes the sum of its inputs and applies a function to this sum. Such a function can be either linear or non-linear (e.g. sigmoid function). A pair of input values (x_1, x_2) are presented to the input layer neurons. After being processed by the input layer, the values are passed to the connections at the hidden layer. They are then modified by applying weights $(w_1, w_2, ...)$ as they pass through this layer. The assigned starting weights are generally random but are then modified during training. The hidden layer neurons process these weighted values. The processed values are passed along the final set of connections and are modified by another set of weights $(u_1, u_{2,...})$ before reaching the output neuron. The output neuron applies an additional process to compute the value of Y_1, the network output. "Bias units" output a fixed unit value analogous to the constant terms in the equations defining the processes carried out by the network.

As noted earlier, "training" is essentially a process of adjusting the connection weights so as to make the network reproduce the known output values in the training set. It is the minimizing of mean square error (between the network and the training sets). Although the initial training can be time-consuming and computer-intensive when large networks are involved, once trained the weights in the network are fixed. Then it is simply a matter of calculating the output value corresponding to any pair of inputs, a relatively quick process. It is the non-linear hidden layer neurons which provide MLPs with their power in modeling data patterns. These units are frequently referred to as "feature detectors" – they decompose the data patterns into simpler features.

However, when we face complicated networks, the precise role of particular hidden neurons is usually very difficult to discern. Although we used a fairly simple and straightforward illustration, neural networks cover a number of algorithm that deal with classification and clustering.

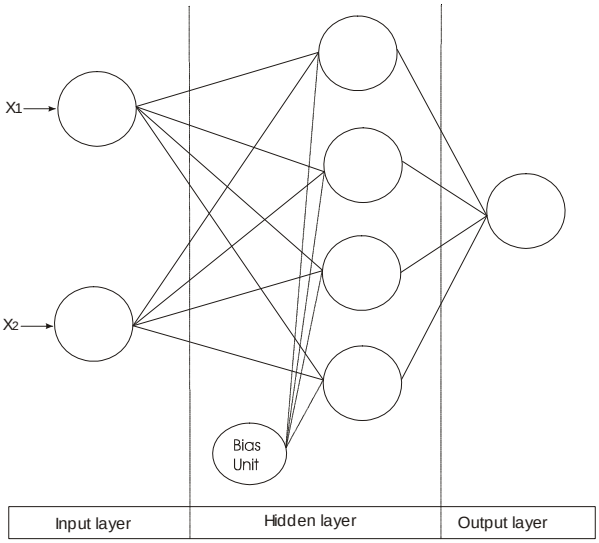

Exhibit 3. Multilayer perceptron: an example

Neural nets have several advantages. They can efficiently combine information from many predictors and cope with correlated independent variables. Compared to traditional techniques such as regression and discriminated analysis, neural nets do handle non-linearities and missing data more effectively. Because of their ability to detect non-linear relationships automatically, neural nets have a significant advantage over regression-type models. Compared to standard multivariate techniques, neural net procedures and results are easier to communicate. Since these models adapt to changing input much more easily than techniques such as multiple regression analysis, they are considered to be especially appropriate in dynamic, fast-changing situations when the relationship between behaviour – (e.g. customer attrition) - and a set of predictors is subject to frequent change.

Neural nets also have several disadvantages. Building the initial neural network model can be very time-consuming since it involves extensive data cleaning, data verification, data transformation, and variable screening. Many of these procedures require specialized skills. Another main limitation of neural nets is that there is no explanation for the outcomes produced by neural nets – I n most cases, it is essentially a black box technique as far as the end-user is concerned. Neural networks need to be “trained”. Training is conceptually similar to deriving weights in a regression equation and involves reading sample data and iteratively adjusting network weights to produce a best prediction. Once such weights are assigned, the model can be applied to others to make predictions. However, training requires large amounts of data. This can be a problem in some cases.

2.6. Genetic algorithms

Genetic algorithms (Gas) are used to solve prediction and classification problems or to develop sets of decision rules similar to the rules that are inferred from decision-tree models. Gas is based on the evolutionary biological processes of selection, reproduction, mutation, and survival of the fittest. They are suitable for use with poorly understood, poorly structured problems because they aim to generate several alternative solutions simultaneously, unlike, say, a regression model which attempts to find a single best solution[1]. Gas can also incorporate in the model any decision criterion. If, for instance, the marketer is interested in maximizing the response rate in a particular segment, this can be built into a GA model (but not into traditional multivariate models such as logistic regression analysis). For example, a GA can explicitly model maximizing the proportion of responses in the top 20% of a direct marketing lift analysis, something logistic regression cannot do. Another feature of Gas is that they are capable of producing unexpected solutions: they may identify combinations of independent variables that may not have been initially obvious. GAs can be used by those who may not be technically skilled. Gas are not suitable for the automatic search of large databases with a large number of candidate variables since GA software tends to be slow because the process of evaluation of the fitness function tends to be time-consuming, when the database and number of variables are large. In such cases decision trees may be more appropriate. Constructing Gas can be quite time-consuming and many runs may be required in the fitting process. GA solutions are difficult to explain as they do not provide statistical measures to enable the user to understand why the technique arrived at a particular solution.

3. The knowledge discovery process

Data mining can be conceived as the knowledge discovery process (KDP). Peter Peacock (2000) provides a model of KDP and the exposition below follows his model (see Exhibit 4). KDP is not new. It is the application of scientific discovery methods to large databases. The terms may be new, but the concepts are not. The elements in Exhibit 4 are described below. Although the exhibit does not have feedback loops, the KDP process is iterative in that there is a substantial flow of information back to prior steps in the process. Although KDP is generally discussed in the context of data mining, it is in fact common to all model building.

[1] Radovic O., Lepojevic V., Genetski algoritam u resavanju NP klase problema", SYM OP IS 1996., Zlatibor, 246-247

Data funneling

Data mining techniques assume that the data quality is high. Data funneling is a set of procedures that ensure that the data collected are suitable for analysis. They include the identification of internal operational data, appropriate external data, moving them to a data repository, evaluating data quality, and obtaining better data when necessary. Data quality is assessed by running simple queries, applying basic visualization techniques, and running automatic validation procedures. The objective here is not to make the data flawless- a near-impossible objective when we deal with large databases-but rather to make sure that there are no gross errors such as wrong type of data, outliers that are clearly wrong and right data in the wrong column.

Data funneling also includes choosing the subset of variables to be analyzed from the larger set of all characteristics available in the data repository.

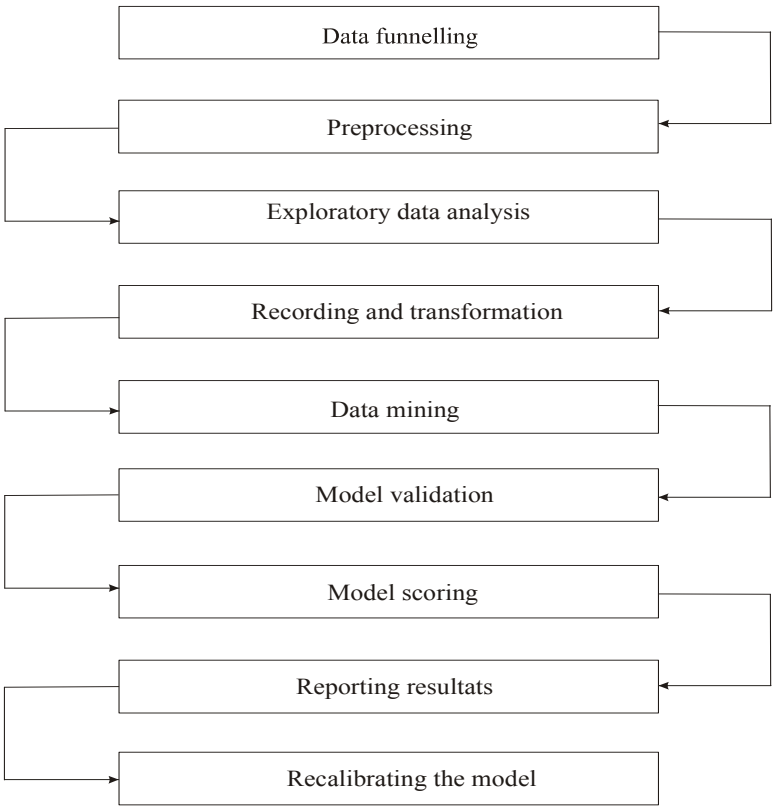

Exhibit 4 The knowledge discovery process

Preprocessing

The next step, data preprocessing, includes the following aspects:

1. *Reformatting*. Formatting data from different sources to a common format.
2. *Standardizing.* Standardizing data attributes to conform to a standard. For example, an organization may have standard specifications for an attribute. The data may have to be converted to conform to this standard. This is particularly true of text-based attributes.
3. *Removing records with sparse data.* Removing records with insufficient information for analysis purposes.
4. *Removing duplicate records.*
5. "*Householding*" When the target unit for the analysis is a household rather than an individual, individuals must be assigned to households. This operation is generally performed by software that looks for sets of common variables such as last names, address components, and phone numbers.

Exploratory data analysis

Exploratory data analysis is used to identify the anomalies and outliers that remain in the data set after it has passed through the previous checks. It also provides the researcher with a "feel" for the preprocessed data through ranges, means, measures of central tendency and dispersion, shape of the distributions, and correlations among variables. The analyst looks for largest and smallest values, central tendency, dispersion, the shapes of the distributions of individual variables, and the structure of the relationships among variables. This step often enables the analyst to generate preliminary hypotheses with regard to the nature of relationships among variables.

Recoding and transformation

In this phase, additional operations are performed on the data. This may include recoding the data to conform to the analyst's hypotheses, creating new variables by combining the existing variables, and the application of data transformation to non-linear data. Data may also be recoded into other values using simple decision rules. Recoding is used to convert continuous data to a nominal form for use with tools such as neural nets and decision trees. It can also be used to convert nominal text label data into numeric values.

Data mining

This phase includes techniques of machine learning from patterns in data using major discovery tools, association rules, decision trees, neural nets, and genetic algorithms.

Model validation

Once the model is built, it is important to assess its validity because models that might have worked on a training set may not work very well when applied to other data. A common approach to model validation is to draw two random samples from the preprocessed data: a "calibration sample" and a "holdout sample" The calibration sample is used to build the model. This model is then tested against the holdout sample to validate the model. If the model performs very poorly on the validation sample then the analyst must modify the model and even rebuild it from scratch.

Model scoring

Model scoring refers to the application of the model developed by the analyst to the entire database. It is done through a set of classification rules developed on the basis of the calibration sample. For example, an equation such as $y = a + b_1 x_1 + b_2 x_2$ developed from a sample of the data is applied to the entire population of records. The scores derived by applying the formula to the entire database, the y s, are placed in a new column in the data base. These are generally known as *scores,* and the process as *scoring.* Scoring may also refer to the process of identifying cluster membership of individual observations when cluster analysis is carried out.

Reporting the results

Once the above processes are completed, the researcher interprets the results and presents them to the decision-maker along with supporting information.

Recalibrating the model

Because marketing is a dynamic process, behaviour patterns identified today may not work two years from now. All models- especially in applied disciplines such as marketing – deteriorate over time, and should be

recalibrated regularly, preferably at definite intervals established beforehand. Recalibrating is the process of rebuilding the model with a recently constructed data set. The recalibrated model may differ from the original I terms of weights applied to the attributes in the models, include new attributes or even have a completely different formulation.

4. Summary

Customer satisfaction research has become commonplace over the last 20 years, with businesses and academic researchers touting continuous improvement strategies driven by customer satisfaction data. Traditionally, researchers have used statistical techniques have limitations, especially in customer satisfaction research. In practice, many researchers ignore assumptions and limitations, which may produce biased and misleading results.

Data mining uses a variety of mathematical algorithms to analyze historical data. The results of this analysis are then used to build models based on real world behavior, which are in turn used to analyze incoming data and make predictions about future behavior.

Data mining can be applied to marketing data in a wide variety of ways. Data mining applications in marketing fall into two broad categories – grouping of variables and identifying functional relationships among variables. While the basic multivariate techniques discussed in this book thus far cover these two aspects, data mining techniques are concerned with achieving these objectives with potentially large databases with a large number of variables. Condensing a large number of variables into potentially meaningful groups and finding relationships among a large number of variables- often without any underlying hypotheses – characterize these techniques.

Literature

1. Akeel Al-Attar, 1998, 'Data Mining – Beyond Algorithms', ttp://www.xpertrule.com/tutor/mining.htm.
2. Berry, J. A. Michael; Linoff, Gordon, 1997, 'Data Mining Techniques: For Marketing, Sales, and Customer Support', John Wiley & Sons, Inc., Canada.
3. Chakrapani C.,2004 , Statistics in Market Research, Arnold, London
4. Garvel Michael S, 2002, Using data mining for customer satisfaction research, Marketing Research, 14(1) 8-12
5. Radovic O., Lepojevic V., 1996, Genetski algoritam u resavanju NP klase problema", SYM OP IS, Zlatibor, 246-247.

KORIŠĆENJE DATA MINING TEHNIKA U ISTRAŽIVANJU TRŽIŠTA

Rezime: Data mining tehnike su od skoro postale popularne kod istrazivača, posebno zbog toga što se njima prevazilaze mnoga ogranicenja koja nameće tradicionalna statistika i što mogu baratati i kompleksnim setovima podataka. Data mining je proces izvlačenja nepoznatih i važnih informacija iz velike baze podataka radi daljeg korišćenja istih za donosenje krucijalnih poslovnih odluka. Data mining tehnike su moćna dopuna klasičnim statistickim tehnikama i nude korisne aplikacije radi istraživanja satisfakcije potrošača.

Ključene reči: Korišćenje podataka, istraživanje tržišta, vizualizacija podataka, pravila pridruživanja, tumačenje bazirano na slučaju, neuronske mreže, genetski algoritam