USE OF DATA MINING TECHNIQUES IN ADVANCE DECISION MAKING PROCESSES IN A LOCAL FIRM

Onur Doğan¹, Hakan Aşan², Ejder Ayçın³
Dokuz Eylül University, Izmir, Turkey

ABSTRACT

In today’s competitive world, organizations need to make the right decisions to prolong their existence. Using non-scientific methods and making emotional decisions gave way to the use of scientific methods in the decision making process in this competitive area. Within this scope, many decision support models are still being developed in order to assist the decision makers and owners of organizations. It is easy to collect massive amount of data for organizations, but generally the problem is using this data to achieve economic advances. There is a critical need for specialization and automation to transform the data into the knowledge in big data sets. Data mining techniques are capable of providing description, estimation, prediction, classification, clustering, and association. Recently, many data mining techniques have been developed in order to find hidden patterns and relations in big data sets. It is important to obtain new correlations, patterns, and trends, which are understandable and useful to the decision makers. There have been many researches and applications focusing on different data mining techniques and methodologies.

In this study, we aim to obtain understandable and applicable results from a large volume of record set that belong to a firm, which is active in the meat processing industry, by using data mining techniques. In the application part, firstly, data cleaning and data integration, which are the first steps of data mining process, are performed on the data in the database. With the aid of data cleaning and data integration, the data set was obtained, which is suitable for data mining. Then, various association rule algorithms were applied to this data set. This analysis revealed that finding unexplored patterns in the set of data would be beneficial for the decision makers of the firm. Finally, many association rules are obtained, which are useful for decision makers of the local firm.

JEL CLASSIFICATION & KEYWORDS
- C80
- D83
- DECISION MAKING
- DATA MINING
- DECISION TREE ALGORITHMS

INTRODUCTION

With the aid of technological tools in organizations, there is a great amount of data being generated by systems. Organizations need to understand large, complex information - rich data sets - in order to become successful in today’s competitive world. In the business world, tools and techniques, which transform data into the knowledge, are becoming recognized as a strategic asset.

Data mining refers to computer-aided pattern discovery of previously unknown interrelationships and recurrences across seemingly unrelated attributes in order to predict actions, behaviors and outcomes. Data mining, in fact, helps to identify patterns and relationships in the data (Frawley, Piatetsky-Shapiro, & Matheus, 1992).

In this study, data mining implementations were achieved in a local firm, which is an actor in the food industry. It can be said that the food industry is generally “information rich,” but unfortunately not all data are mined, which is required for effective decisions. One of the major problems facing food industry managers/decision makers is determining the right price at the right time. Poor decisions can lead to damaging consequences which are unacceptable for food industry. Main purpose of this study to give information to the decision makers about price of products that they haven’t recognized yet.

Data Mining

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

Data mining is the process of searching and analyzing data in order to find implicit, but potentially useful, information (Frawley et al., 1992). It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases (Berry & Linoff, 1997). To undertake these large data analysis projects, researchers and practitioners have developed new methods targeted at large data mining problems (Hand, Mannila, & Smyth, 2001; Zhang & Zhou, 2004).

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization (Han & Kamber, 2001, p. 1)

Data mining methods are divided into two categories: descriptive and predictive. Descriptive methods concerned with answering “what is?” and predictive methods concerned with finding out “what will be”. Clustering, association rule discovery, etc. counted among descriptive methods and sequential pattern discovery, classification, regression, etc. counted among predictive methods.

A decision tree is a prediction model and an algorithm that classifies the data inputs by using the branches and leaves of a tree. This tree tries to predict the outcomes based on preexisting database information comprised of exact inputs and their outputs (Kabra & Cichkar, 2011, p. 11). The learning and classification steps of decision tree induction are simple and fast, and decision tree classifiers have good accuracy. Decision trees are the suitable technique for the
aim of this study, so it will be shortly described in the following parts.

**Decision Tree Algorithms**

Decision trees have obvious value as both predictive and descriptive models. Prediction can be done on a case-by-case basis by navigating the tree. More often, prediction is accomplished by processing new cases through the tree or rule set automatically and generating an output file with the predicted value or class appended to the record for each case (Bounsaythip & Esa, 2001:20).

Decision tree algorithms implemented in literature include, Hunt’s Concept Learning System (Hunt, Marin, & Stone, 1966), Quinlan’s ID3 (Quinlan, 1986), ACLS (Patterson & Niblett, 1983), Classification and Regression Trees (CART), Kass’ Chi-squared Automatic Interaction Detection (CHAID) (Kass, 1980), Quinlan’s C4.5 and C5.0, SLIQ (Mehta, Agrawal, & Rissanen, 1996) and SPRINT (Shafer & Agrawal, 1996). Table 1 explains the decision tree algorithms and their characteristics.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS</td>
<td>Constructs a decision tree that attempts to minimize the cost of classifying an object. Uses a look ahead strategy similar to minimax. At each stage, CLS explores the space of possible decision trees to a fixed depth, chooses an action to minimize cost in this limited space, then moves one level down in the tree. Depending on the depth of look ahead chosen, CLS can require a substantial amount of computation, but has been able to unearth subtle patterns in the objects shown to it.</td>
</tr>
<tr>
<td>ID3</td>
<td>Generates all possible decision trees that correctly classify the training set and to select the simplest of them. The number of such trees is finite but very large, so this approach would only be feasible for small induction tasks.</td>
</tr>
<tr>
<td>ACLS</td>
<td>Generalization of ID3. CLS and ID3 both require that each property used to describe objects has only values from a specified set. In addition to properties of this type, ACLS permits properties that have unrestricted integer values. The capacity to deal with attributes of this kind has allowed ACLS to be applied to difficult tasks such as image recognition.</td>
</tr>
<tr>
<td>CART</td>
<td>Binary split based GINI (recursive partitioning motivated by statistical prediction), exactly two branches exist from each non-terminal node. Pruning based on measure of complexity of the tree. Supports classification and regression. Handles continuous target variables. Requires data preparation.</td>
</tr>
<tr>
<td>CHAID</td>
<td>Multi-way splits using chi-square tests. The number of branches varies from two to the number of predictor categories.</td>
</tr>
<tr>
<td>C 4.5 and C 5.0</td>
<td>Produce tree with multiple branches per node. The number of branches is equal to the number of categories of predictor. Combine multiple decision trees into a single classifier. Use information gain for splitting. Prune based error rate at each leaf.</td>
</tr>
<tr>
<td>SPRINT</td>
<td>For large dataset. Splitting based on the value of a single attribute. Removes all memory restrictions by using attribute list data structure.</td>
</tr>
</tbody>
</table>

Table 1: Decision Tree Algorithms

C5.0 is another new decision tree algorithm developed based on C4.5 by Quinlan (1996). It includes all functionalities of C4.5 and apply a group of new technologies, among them the most important improvement is “boosting” technology for building up the accuracy rate of identification on samples.

Boosting algorithm sets weight for each sample, which presents its importance. The higher the weight, the more the sample influence on the decision tree. Initially, every sample has the same weight. In each trial, a new decision tree is constructed. The weight of each sample is adjusted, such that the learner focus on the samples which are misclassified by the decision tree constructed in the last trial, which means these samples will have higher weight (Quinlan, 1996; Freund & Shapire, 1997).

Another difference between C4.5 and C5.0 is the construction of a cost-sensitive tree. In C4.5 all kinds of misclassification errors are treated equally and the object of decision tree minimizes the number of misclassification errors. But many practical classification problems have different costs associated with different types of error. The cost of misclassification presents the seriousness of some misclassification errors. Different costs of misclassification can be showed in cost matrix. Cost matrix shows the different combinations of the predicted class and actual class. In C5.0, misclassification cost can be set in costs matrix by a certain value. By minimizing expectation of the total misclassification costs, C5.0 constructs a decision tree called cost-sensitive tree (Pang & Gong, 2009:96).

**Case Study**

In this part, data set structure, data preparation and implementation parts will be given. As we mentioned before; application was performed by taking data from a company which operates in the food sector in Turkey. Aim of these analyses were increasing sale of the company and providing the sale of correct product with correct price. For this study, approximately 50,000 sales data were extracted from the database of the company.

There are several fields in data set that we can include to analysis (price, payment information, import and export information, etc.), listed as follows;

- **Product Name:** The product is being sold by the company
- **Product Category-1:** This category indicates product’s main type (meat or side product)
- **Product Category-2:** This category indicates which product belongs to which animals (Calf, sheep, etc)
- **Product Category-3:** This category shows that product is boned or boneless
- **Product Category-4:** This category shows the company’s special reporting codes
- **Customer Name:** Customer who is bought product from the company
- **Customer Category-1:** This category shows type of sale (retail, wholesale or distributor)
- **Customer Category-2:** This category indicates type of customer (chain store, hotel, distributor)
- **Customer Category-3** This category indicates if the customer private enterprise or public enterprise
- **Customer Category-4:** This category indicates in which region the user customer occupy ( İzmir, İstanbul, etc.)
- **Quantity:** Quantity of sold product
- **Price:** Sale price of sold product
- **Project Code:** This indicates which project be

When examining the data on the system, it has been observed that some data were input for trial purposes or incorrectly by users. While investigating the data, it has been...
noticed that some variables were inputted as ‘0’ or missing. These data were deleted from the system.

Product name and customer name category have been excluded from the data set, because these fields have vast amount variables, which could darken the decision tree. The field, ‘price of the product’ could take many different value which is not suitable for the technique that will be used. This field cause so many branches in the decision tree which makes difficult to understand the structure of the tree. For this reason data categorization is applied with the aim of decreasing the multivariability. It has been decided that (by consulting the managers of the firm) the field should be categorized 6 different categories. This categorization was applied by the help of IBM SPSS, visual binning menu and categorized data were used for the analysis.

Customer category 1 indicates type of sales as retail, wholesale or distributor. Most of the data is belong to distributors, so this field canceled and only distributors’ data have been included the analysis. Similarly, Customer Category-3 also canceled since most of the data have belong to private enterprises. Izmir region has been chosen to investigate and just this region’s data were taken the data set. So, Customer Category-4 also has been excluded from the data set.

After the corrections that mentioned, the dataset size has reduced to 19,676 and the data set has 8 fields (see Table 2).

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Category-2</td>
<td>Set (Nominal)</td>
<td>In</td>
</tr>
<tr>
<td>Product Category-1</td>
<td>Flag (Binary)</td>
<td>In</td>
</tr>
<tr>
<td>Product Category-2</td>
<td>Set (Nominal)</td>
<td>In</td>
</tr>
<tr>
<td>Product Category-3</td>
<td>Set (Nominal)</td>
<td>In</td>
</tr>
<tr>
<td>Product Category-4</td>
<td>Set (Nominal)</td>
<td>In</td>
</tr>
<tr>
<td>Quantity</td>
<td>Range (Scale)</td>
<td>In</td>
</tr>
<tr>
<td>Project Code</td>
<td>Set (Nominal)</td>
<td>In</td>
</tr>
<tr>
<td>Price (Binned)</td>
<td>Set (Nominal)</td>
<td>Out</td>
</tr>
</tbody>
</table>

Once the data set preparation is complete, the data are transferred to SPSS Clementine. Figure 1 shows the screen of the fields, types, values, and directions.

In the next step, input variables and output variables were specified. Because the required variable is the price, this field was set as an output variable.

| Table 3: Performance of the Different Test Options |
|-------------------|-------------------|
| Method            | Accuracy          |
| 10-fold cross validation | 62.24%            |
| Percentage Split (66% for training) | 55.10%            |
| Percentage Split (80% for training) | 58.10%            |
| Percentage Split (90% for training) | 60.12%            |

Many different classification algorithms have been implemented the data set. Some of them has given very poor results and some of them fairly well results. But most accurate and useful results obtained from the C5.0 algorithm. And, there are also many ways to choose training and test data set from the main data set. One of the common methods when choosing training set is taking 2/3 of the data set as a training set and remaining 1/3 of the data as a test data. K-fold cross validation is another way to estimate the performance of a classifier. According to this method, researcher divide the data into k folds and take the k-1 folds as a training data set and 1 fold for the test. After that by repeating this process k times and calculating the mean of the accuracies, final performance score is obtained. Besides this; according to the data size, data structure, etc. researchers could make a percentage split on the data set and could create training set and test set (i.e. 80% of the data for training and 20% of the data for test). Table 3 shows some accuracy values that were obtained from different performance evaluation methods.
CONCLUSION

In this study, data mining tools and techniques have been implemented the data set, which belongs to a firm that is an actor in the food industry in Turkey. As we mentioned before C5.0 algorithm and k-fold cross validation performance estimator have been used during the mining. The model gives some useful results. Some of the results that have been extracted from the model which will help to decision maker for their decisions are as follows,

- It is obtained that although there is a lot of important variables to determine the price, the most important variable is Product Category-2.
- Other important categories, which influence determination of project code, are product quantity, etc.
- Project code is important for pricing packed products as block.
- The amount is an important identifier. It is found in almost all of the branches.

The results from the data mining process of the food firm’s records showed us that there is still some useful knowledge lying in database which could be really useful for the decision makers.

By giving decision by taking into consideration of data mining results could improve the efficiency and quality of services given to clients.

Data mining is a useful field to advanced decision processes. But, some small disadvantages still occur on data mining processes. For example; there are a lot of decisions (which algorithm will be used, which performance estimator will be used, which attributes will be included the model, etc.) should be made by the implementers. Besides this, the results that obtained from any data mining processes still need to human interference. Nevertheless, using data mining tools and techniques for decision is increasingly going on and in some cases these tools and techniques have no other alternative.

REFERENCES


