

USAGE OF DATA MINING TECHNIQUES IN DISCOVERING THE FOOD CONSUMPTION PATTERNS OF STUDENTS AND EMPLOYEES OF UNIVERSITY

Ahmet Selman BOZKIR¹

Ebru SEZER²

^{1,2} Hacettepe University Computer Science and Engineering Department

¹e-mail: selman@cs.hacettepe.edu.tr

² e-mail: esezer@cs.hacettepe.edu.tr

ABSTRACT

In this study, the data, which include daily food lists and sales records of 2008 that are belonged to refectories of Hacettepe University at Beytepe Campus, are analyzed via two data mining techniques (decision trees and association rules). By this way, the patterns and important factors that affect the amount of consumption are tried to be determined. With the help of the decision tree models and analysis, an estimated consumption prediction success almost 80% is achieved.

1. INTRODUCTION

Catering is the act of providing food and services or it may be defined as preparing or providing food for someone else to serve; or preparing, delivering and serving food at the premises of another person or event [1]. Catering factories and refectories of public institutions present different types of menus to their employees and consumers every day. While in some institutions the demand for the food is stable, in many large institutions such as universities, hospitals, etc. the demand of food may vary. One of the important reasons of this alteration is the variation in the interest of people to the menus that are presented.

Actually, lots of catering factories and refectories don't apply an optimal way in production. Demand for a menu in a day, may be less than expected value or vice versa. Moreover, the demand for a menu that's not presented yet; can't be predicted correctly. The future prediction in demand for a menu is based on human prediction performance in today's catering services. Therefore, some problems may occur in this scenario. Furthermore, the current approach is unproductive in economics point of view.

Data mining is a process that uses a variety of data analysis tools to discover patterns and relations in data that may be used to make predictions. Supervised data mining techniques are used to model an output variable based on one or more input variables and these models can be used to predict or forecast future cases [2].

In the cases which the number of people who consume food is not recognized in certain, prediction of demand for the menus in advance will make valuable contribution to this issue in terms of labor force and productivity.

In this paper, the data, which include daily food lists and sales records of 2008 that are belonged to refectories of Hacettepe University at Beytepe Campus, are analyzed via two decision trees and association rules data mining techniques. Some important and frequent patterns are discovered and presented in this study.

In a related work, time series, a data mining technique, is employed to forecast daily demand of perishable ingredient for a worldwide fast-food restaurant. In addition, they illustrated how Box-Jenkins seasonal ARIMA time series models are used to discover outliers in demand [7].

2. DATA MINING

Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules [3]. In other words, data mining is the overall process of extracting meaningful patterns and relationships in data by using methods like artificial intelligence, machine learning and statistics via advanced data analysis tools. Oracle BI, SPSS Clementine, SAS Enterprise Miner and Microsoft Analysis Services are well known data mining tools in marketplace [4]. Data mining methods are classified into two categories as follows:

- Predictive methods (decision trees, Bayes classifiers, rule based classifiers etc...)
- Descriptive methods (clustering, association rules analysis etc...)

The goal of predictive methods is making predictions on new cases by using old cases. However, the aim of descriptive methods is discovering hidden relationships, correlations and descriptive attributes in current data.

In this study, both of these method groups are utilized. SQL Server 2008 Analysis Services data mining tool is selected as the analysis platform of this study. We used "Microsoft Decision Trees" method in making future predictions on demand of presented and non-presented menus. Likewise, association rules method is employed in discovering frequent consumption habits of people. Furthermore, the dependency graph utility of this software package is utilized to reveal the factors that affect food consumption habits of students and three different kinds of employees.

2.1. DECISION TREES

The decision tree is probably the most popular data mining technique. The most common data mining task for a decision tree is classification; for example, to identify the credit risk for each customer [5]. As it has a good visual representation and ease of understanding decision tree is well used method in classification task.

The principle idea of a decision tree is to split the data recursively into subsets so that each subset contains more or less homogeneous states of the target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on predictable attribute. When this recursive process is completed, a decision tree is formed [5]. Unknown cases can be classified after building decision tree. While discrete valued trees are

called decision trees, numeric valued trees are called regression trees and they are used in predicting numerical values. ID3, C4.5 and CART (Classification and Regression Trees) are well known decision tree algorithms [5]. In this study, Microsoft Decision Tree algorithm is used in decision tree analysis stage. The main reason of this selection is that it builds dependency network graph which shows the relationships and effect of variables on predictable attribute(s).

2.2. ASSOCIATION RULES MINING

Association rule mining is one of the most well studied mining methods. Such rules associate one or more attributes of a dataset with another attribute, producing an if-then statement concerning attribute values. Mining association rules between sets of items in large databases was first stated by Agrawal, Imelinski and Swami in 1993 and it opened brand new family algorithms [6]. Apriori algorithm is probably the most used algorithm in association rules mining. In Apriori algorithm, the support and confidence parameters influence the quality and quantity of extracted rules from data.

In this study, the frequent associations in food consumption habits are tried to be discovered via association rules mining technique.

3. MATERIALS AND METHODS

At the startup stage of this research, daily food sales records data is extracted from the turnstile machines of Hacettepe Beytepe Campus refectories. The data consist of both the lunch and dinner sales numbers of 3 different types of employees and students. The daily menu including four diverse types of foods are also linked to this dataset on daily basis. In addition, day and month numbers and a flag variable which is denoting the day is a holiday or weekend are also added to dataset for better analysis and to check whether day and month variables and also holidays have any effect on food consumption habits of people. Finally, dataset which contains the attributes given in Table 1 is obtained; in addition, it has records of 365 days.

Table 1: Attribute properties of final dataset.

Attribute name	Type	Usage	Description
Day	Number	Input	Day number varying from 1 to 31
Month	Number	Input	Month number varying from 1 to 12
Day name	Text	Input	Day names starting by Monday
IsHoliday	Boolean	Input	0 or 1, denoting that day is a holiday or not
Food 1	Text	Input	First food name in menu
Food 2	Text	Input	Second food name in menu
Food 3	Text	Input	Third food name in menu
Food 4	Text	Input	Fourth food name in menu
Lunch-Academics	Number	Predict	Sales numbers of academics at lunch
Dinner-Academics	Number	Predict	'' '' of academics at dinner
Lunch-Officials	Number	Predict	'' '' of officials at lunch
Dinner-Officials	Number	Predict	'' '' of officials at dinner
Lunch-Contractual	Number	Predict	'' '' of contractual workers at lunch
Dinner-Contractual	Number	Predict	'' '' of contractual workers at dinner
Lunch-Students	Number	Predict	'' '' of students at lunch
Dinner-Students	Number	Predict	'' '' of students at lunch

After the data collection, data cleaning and preprocessing steps are carried out via normalization of food names and fixing some typographic mistakes on data. Resulted clean dataset is then exported to Microsoft SQL Server database from Excel sheet. Once the database is ready in SQL Server 2008, Analysis Services, the analyzing platform, can easily make analysis on dataset.

As the purpose of this research is discovering the factors which affect food consumption numbers and habits of people together with making daily consumption predictions, a decision tree model is created with Microsoft Decision Tree algorithm. The parameters of algorithm are set with default settings. In the analysis, some attributes such as food names are used as input variables whereas some which are exemplified as student-dinner are used as predictable variables. The factors which affect the 3 types of employees and students lunch & dinner consumptions are extracted from dependency network graph shown in Fig. 1.

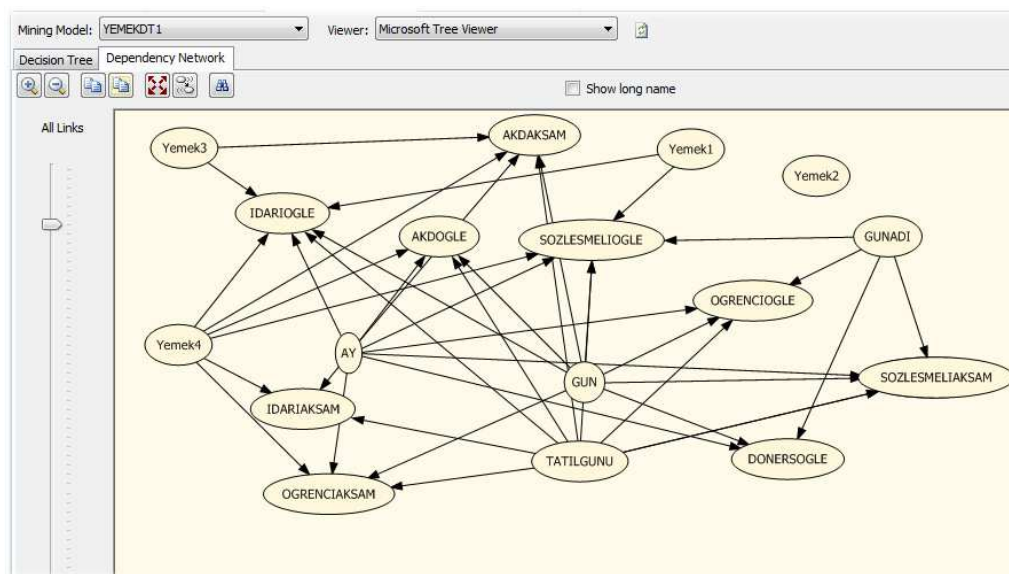


Figure 1: The dependency network graph obtained from decision tree model

In association rules mining stage, the frequent associations are searched in data. As the whole dataset has very limited cases such as 365 records, the support and confidence parameters of Apriori algorithm in Analysis Services are set as 2 and %20 respectively and the maximum item set count in a rule is set as 4 to reveal remarkable rules.

4. RESULTS AND DISCUSSIONS

In this study, it is aimed to discover important factors that affect food consumption, build a platform to predict daily consumption in the combination of four diverse types of foods and discover frequent associations in consumption. The first goal is accomplished with the help of dependency network graph in decision tree analysis. As shown in Table 2, first four factors that affect consumptions on different types of people are listed with their percentage values.

To achieve second goal, a prediction query run against original data in Analysis Services and results are compared with original consumptions. Prediction performances are measured in R^2 unit. Results of student-lunch and officials-lunch predictions are shown in Fig. 2.

Table 2: Predictable attributes and the factors that affect them

Target Attribute	1.Factor	2.Factor	3.Factor	4.Factor
Student-Lunch	IsHoliday (80%)	Month (68%)	Day (65%)	-
Student-Dinner	Month (70%)	Day (67%)	IsHoliday (50%)	Food 4 (25%)
Academics-Lunch	IsHoliday (89%)	Month (86%)	Day (85%)	Food 4 (46%)
Academics-Dinner	Month (96%)	IsHoliday (52%)	Food 3 (29%)	Day (25%)
Officials-Lunch	IsHoliday (85%)	Day (83%)	Month (82%)	Food 3 (52%)
Officials-Dinner	IsHoliday (53%)	Month (44%)	Food 4 (30%)	Day (10%)
Contactuals-Lunch	Month (100%)	Day (96%)	IsHoliday (75%)	Food 1 (17%)
Contactuals-Dinner	IsHoliday (51%)	-	-	-

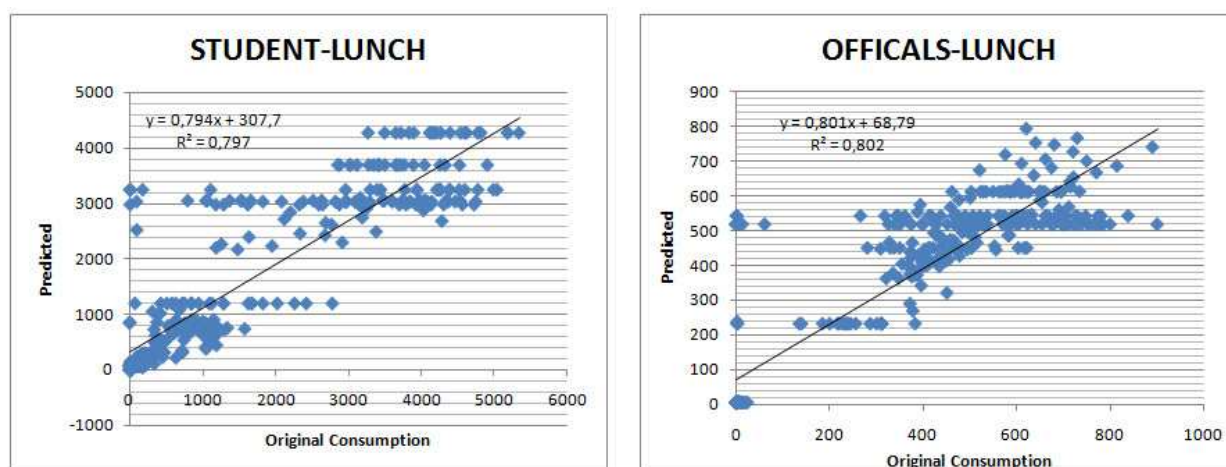


Figure 2: Some predicted consumption values versus original data

As the prediction performances are measured in R^2 , the overall prediction performance results are listed in Table 3.

Table 3: Demand prediction performances of three types of employees and students

The Community	Type of Meal	R^2 Value
Student	Lunch	0,797
Student	Dinner	0,709
Students	Lunch + Dinner (Total)	0,780
Officials	Lunch	0,802
Academics	Lunch	0,774
Contactuals	Lunch	0,853

In association rules mining research of this study, associated habits of consumers are tried to be discovered via Apriori algorithm. As the training dataset has very limited numbers of cases, the support and confidence values are set slightly low such as 2 and %20 respectively in this study. Some interesting rules are listed in Table 4.

As the confidence values may reflect some inaccurate results and to be more precise, lift values are used to evaluate the interestingness of the rules.

Table 4: Some interesting rules that derived from association rules mining study

Support	Lift (Importance) Value	Rule
2	0,68	Food3 = Barbunya Pilaki, Food4 = Revani -> Academics-Lunch \geq 1223
4	0,22	Food2 = Etli Pırasa, Student-Lunch $<$ 588 -> Student-Dinner $<$ 284
3	1.103	Food2 = Pilavüstü Et Döner, Food3 = Ayran -> Student-Total \geq 5553

5. CONCLUSIONS

In this study, a platform which can be used in optimization of the meet for demand in food consumptions is built by using two different data mining algorithms. Also the factors which may affect the consumptions are tried to be discovered. The major beneficiaries of this platform would be catering firms and refectories of public institutions in the situation that the consumption amounts are not certain.

To build such a platform, 2008 consumption dataset of Hacettepe refectories are used as the basis of this study. Although the overall (365 days) prediction performances are fine, some outliers (inaccurate predictions) can be seen in Figure 2. To overcome this problem and to be more precise in daily predictions it's obvious that much more training cases (more than 1 year) and some feasible attributes (ex: calories of a menu) is required in training dataset.

As seen in this study, data mining methods can be used as a decision support system in resource optimization problems of food sector.

6. REFERENCES

- [1] Kahraman, C., Cebeci, U. and D. Ruan, 2003, "Multi-attribute comparison of catering service companies using fuzzy AHP: The case Turkey", *International Journal of Production Economics*, Vol. 87, pp. 171-184.
- [2] Hearty, P.A. and M.J. Gibney, 2008, "Analysis of meal patterns with use of supervised data mining techniques – artificial neural networks and decision trees", *The American Journal of Clinical Nutrition*, Vol. 88, pp. 1632-1642.
- [3] Berry, M. and G. Linoff, 2000, *Mastering data mining: The art and science of customer relationship management*, John Wiley & Sons.
- [4] Bozkir, A.S., Gök, B. and E. Sezer, 2008, "İnternetin eğitimsel amaçlar için kullanımını etkileyen faktörlerin veri madenciliği yöntemleriyle tespiti", *Bilimde Modern Yöntemler Sempozyumu*, Eskişehir, Turkey.
- [5] Tang, Z. and J. MacLennan, 2005, *Data Mining with Sql Server 2005*, John Wiley & Sons, Indiana.
- [6] Romero, C. and S. Ventura, 2007, "Educational data mining: A survey from 1995 to 2005", *Expert System with Applications*, Vol. 33, pp. 135-145.
- [7] Liu, L., Bhattacharyya, S., Sclove, S.L., Chen, R. and W.J. Lattyak, 2001, "Data mining on time series: an illustration using fast-food restaurant franchise data", *Computational Statistics & Data Analysis*, Vol. 37, pp. 455-476