ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 759–763

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT-2010

Predicting food demand in food courts by decision tree approaches

Ahmet Selman Bozkir [a] *, Ebru Akcapinar Sezer [a]

[a] Hacettepe University Computer Engineering Department, Beytepe, Ankara 06532,Turkey

Abstract

Fluctuations and unpredictability in food demand generally cause problems in economic point of view in public food courts. In this study, to overcome this problem and predict actual consumption demand for a specified menu in a selected date, three decision tree methods (CART, CHAID and Microsoft Decision Trees) are utilized. A two year period dataset which is gathered from food courts of Hacettepe University in Turkey is used during the analyses. As a result, prediction accuracies up to 0.83 in R^2 are achieved. By this study, it's shown that decision tree methodology is suitable for food consumption prediction.

Selection and/or peer-review under responsibility of the Guest Editor.

Keywords: Decision Trees, Food Demand, CART, CHAID, Microsoft Decision Trees

1. Introduction

Catering is defined as the act of providing food and services or concretely 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] . In today's world, catering services are generally presented in food courts, restaurants, and fast food companies. While the demand for food remains stable or constant in many occasions, in some constitutions (i.e., universities, federal buildings and private organizations) variations can be observed because of several reasons. Seasonal demand changes, holidays, charm of the menus can be counted as examples of these reasons. On the other hand, these alterations cause problems in economic point of view. Private companies and other institutions which have their own food courts, lose money and resources in the case that supply is more than demand. Besides, in some days due to these unforeseen changes, the amount of produced food can't be adequate to meet demand. This situation is also causes other problems such as lose of customers and prestige. Existence of a system or technique which predicts the demand for the amount food for a given menu for a specific date could benefit to the institutions in many ways as it optimizes the balance between supply and the demand for saving resources. In essence, this is a real resource optimization problem and data mining techniques can be successfully applied in solving this kind of problems.

Data mining, on the other hand, is discovering hidden relationships, correlations, trends and associations in data with the help of smart algorithms over automatic or semi-automatic approaches. Besides, data mining involves various techniques such as statistics, neural networks, decision tree, genetic algorithm, and visualization techniques that have been developed over the years [2]. In general, data mining methodologies are classified as predictive and descriptive methods. While the purpose of descriptive methods is defining the nature of data, making future predictions based on current observations is the main goal of predictive ones. Therefore, predictive data mining approaches are suitable solution candidates for the main problem defined in this study.

Data mining and similar artificial intelligence (AI) methodologies like neuro-fuzzy are the methodologies which the researches have used in resource planning or optimization for many years. Literature has many examples of this

1877-0509 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.procs.2010.12.125

kind of studies. For instance, Abiyev et al. [3] investigated the applicability of neural network based fuzzy inference systems (neuro-fuzzy) for electricity consumption prediction and used their developed system for prediction of future values of electricity consumption in Northern Cyprus. Yurdusev et al. [4] employed generalized regression neural networks (GRNN) technique for municipal water consumption prediction. Similarly, Altunkaynak et al. [5] used Takagi-Sugeno (TS) fuzzy logic method for predicting future monthly water consumption values in Istanbul city. On the other hand, Chen and Wang [11] used support vector regression method in tourist demand prediction by enriching it with genetic algorithms. However, the usage of data mining and related techniques in the food consumption field is very limited and the number of studies about this subject is considerably low. Likewise, Bozkir and Sezer [1] applied Microsoft Decision Trees algorithm for revealing food consumption patterns in a refectory of a university. Besides, Bhattarcharyya et al. [6] employed time series, a well-known predictive data mining technique, method to forecast daily demand of perishable ingredient for a worldwide fast-food restraunt. Furthermore they showed how Box-Jenkins seasonal ARIMA time series models can be used to reveal outliers in demand.

In this study, three decision tree algorithms (CART, CHAID and Microsoft Decision Trees) are employed to predict menu based food consumption demand in food courts of Hacettepe University (Turkey) for various customer types. The reasons behind the selection of decision trees as the basis method of this study can be listed like this: (1) decision tree models are easily understandable and interpretable models so decision makers can make decisions even lack of business analyst or data mining expert (2) decision tree models are quick to build and on the fly predictive methods therefore it’s highly suitable and adaptable for upcoming data changes as it has low training time [9]. Moreover, decision trees techniques have high prediction accuracy in many fields so that makes them preferable and trustable choices for this kind of task. With the help of these methods, daily consumption patterns are revealed and a generally accurate predictor is obtained for menu based food consumption prediction. In this study, R^2 (determinant of coefficient) performance indicator is considered as the measure of prediction performance and highly accurate results like 0.83 are achieved.

2. Data Mining & Decision Trees

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 [8]. In other words, data mining is the whole process of discovering meaningful patterns and relationships in data by using methods like artificial intelligence, machine learning and statistics via sophisticated data analysis tools. On the other hand, data mining methods are grouped as predictive methods and descriptive methods. Predictive methods such as decision trees, Bayes classifiers, support vector machines etc. are aimed to make future predictions on unseen cases based on past observed cases. Predictive approaches are often called supervised approaches due to its training phase requirement. On the other hand, descriptive data mining approaches like clustering are aimed to reveal hidden relationships or correlations in data via unsupervised techniques.

In practice, there are several data mining tools such as Oracle DM, Microsoft Analysis Services, SPSS Clementine, and SAS Enterprise Miner for commercial use [2]. In this study, SPSS Clementine and Microsoft Analysis Services are employed in whole steps covering data preparation to report generation. Microsoft Analysis Services is preferred as it owns Microsoft Decision Trees algorithm and having support for programmability via AMO and ADOMD extensions that can be integrated with .NET languages like C# or Visual Basic. On the other hand, SPSS Clementine is selected due to its native support for CHAID (Chi Squared Automatic Interaction Detection) and CART (Classification and Regression Trees) algorithms.

As stated before, decision tree is a data mining approach generally used for prediction and classification. Although other methodologies such as neural networks and rule based classifiers are the other options for classification, decision tree has the advantages of interpretation and understanding for the decision makers to compare with their domain knowledge for validation and justify their decisions [10]. Some decision tree algorithms are not designed for only classification but also regression. While every decision tree algorithm has support for discrete value prediction (classification), some of them have also support for prediction on continuous valued attributes. The general idea of a decision tree is splitting the data recursively into subsets so that each subset contains more or less homogeneous states of target predictable attribute. At each branch in the tree, all available input attributes are calculated again for their own impact on the predictable attribute [9]. When this recursive process is completed, the pruning step is started if the algorithm supports. After all, final decision tree is formed. ID3, CART, CHAID, QUEST, C4.5 and Microsoft Decision Trees are the famous and well-known decision tree algorithms. In next sections, readers can find brief information about decision tree methods employed in this study.

2.1. A brief overview of CART, CHAID and Microsoft Decision Trees algorithms

Classification and Regression Tree (CART) algorithm is first proposed by Breiman et al. [12] for both classification and regression tasks in 1984. CART algorithm uses Gini index measure as the splitting criteria and does only binary splitting. Moreover it provides a tree pruning stage for avoiding over fitting problems. As its strategy is based on two-way splitting, it is generally not preferred for the decision tree analyses which are focused on investigation of variable distribution over tree nodes (i.e. finance sector). However this novel approach makes CART a powerful prediction tool which yields generally accurate results.

CHAID algorithm was developed by J. A. Hartigan from a method called Automatic Interaction Detection. As it uses Chi-squared test for tree splitting strategy it is called Chi Squared Automatic Interaction Detection. CHAID algorithm supports continuous and discrete valued variables as input like CART and can do regression and classification tasks over outcome variable. Moreover, it does not serve a tree pruning stage. As stated before, CHAID uses multi-way splitting strategy thus it creates more interpretable models than CART for decision makers.

Microsoft Decision Trees (MSDT) is an algorithm that is shipped with Microsoft Analysis Services product package for classification and regression tasks. MSDT algorithm mainly employs Shannon’s entropy in tree splitting and has support for analyzing discrete and continuous valued attributes. However, like CHAID, it does not have any pruning step but presents a parameter called “complexity penalty” for controlling tree growth. MSDT serves both multi-way and binary splitting options to data miners. Furthermore, MSDT has also automatic feature selection and cardinality reduction features that are not available in other data mining packages [9].

3. Data

In this study, the dataset which has two years period of daily menus and the sales numbers for variety of customer types belonging to Hacettepe University's food courts is employed during the training and test phases. The variables in dataset and their corresponding attributes are listed in Table 1.

Table 1. The variables and their attributes

Variable name	Type	Usage	Description
Day	Continuous	Input	Day number ranging from 1 to 31
Month	Continuous	Input	Month number ranging from 1 to 12
Day Name	Discrete	Input	The name of the day ranging from Monday to Sunday
Is Holiday	Boolean	Input	A flag variable (0/1) denoting that day is in weekend/holiday or weekday
Calorie	Continuous	Input	Total calorie amount of the menu
Food1	Discrete	Input	First food name in menu
Food2	Discrete	Input	Second food name in menu
Food3	Discrete	Input	Third food name in menu
Food4	Discrete	Input	Fourth food name in menu
Sales-Student-Lunch	Continuous	Predict	Number of sales for students in lunch session
Sales-Students-Dinner	Continuous	Predict	Number of sales for students in dinner session
Sales-Academic-Lunch	Continuous	Predict	Number of sales for academic staff in lunch session
Sales-Officials-Lunch	Continuous	Predict	Number of sales for officers in lunch session
Sales-Contractual-Lunch	Continuous	Predict	Number of sales for contractual employees in lunch session

Customers are counted by turnstile machines in selected university. Thus, every person increments the sales counter one by one by according to his/her type. In this study, total daily sales of students, academic staff, officers and contractual employees are considered with in lunch and dinner sessions. However, due to lack of adequate numbers of sales in dinner sessions of academic staff, contractual employees and officers, only the dinner session of students is included in dinner analysis. Although, there are 730 cases for two years period, due to null values of calorie attribute in 103 days, these 103 cases are removed in target dataset. Besides, some food names which are different but pointing same food are unified to single name for better prediction accuracy and generalization.

4. Applying Decision Tree Methods

After the preprocessing stage, 627 days are randomly partitioned into a training set of 533 days (%85) which is used to develop models and a test dataset of 94 days (%15) used for model assessment. During model building stages, SPSS Clementine 12 is employed for CART and CHAID studies (see Fig. 1). On the other hand, Microsoft Analysis Services is used in MSDT study. For simplicity and prevention from over-fitting, depth of CART and CHAID trees is limited to maximum 7 depths. However, due to lack of "depth of tree" setting in MSDT, "complexity-penalty" parameter is kept with its default value in algorithm. Furthermore, multi-way splitting approach is used in MSDT training phase. Then, test dataset is tested against these three decision tree models and R^2 performances are measured. Gained R^2 performances are listed below in Table 2.

Table 2. The R^2 performances of various decision tree models

Customer Type	MSDT	CART	CHAID
Student-Lunch Session	0.800	0.729	0.800
Student-Dinner Session	0.585	0.549	0.525
Officers-Lunch Session	0.627	0.673	0.721
Academic Staff-Lunch Session	0.616	0.506	0.595
Contractual Employees-Lunch Session	0.722	0.741	0.838
Average:	0.670	0.639	0.695

As can be easily seen on average values, CHAID is the leading algorithm with most accurate prediction results. On the other hand with a slight difference, MSDT comes secondly and finally CART becomes the third. Comparatively, only the models of student-lunch and contractual employees-lunch sessions perform adequate prediction performances.

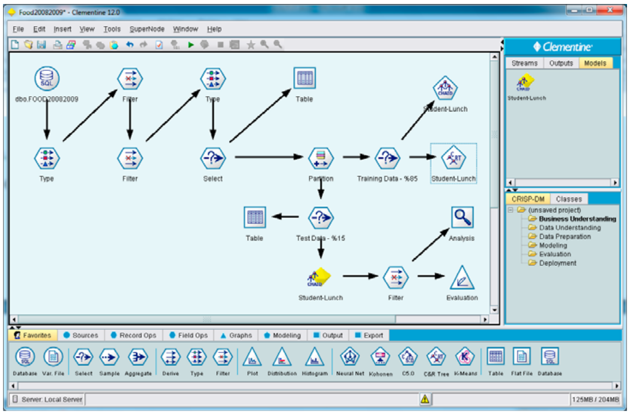

Figure 1. Decision tree modeling environment in SPSS Clementine 12

5. Conclusions and future work

In this study, two main goals are targeted: (1) building data mining models for demand prediction of food consumption; (2) identifying the factors affecting the consumptions for every type of customers. Although two of these aims are achieved, due to fact that this paper has limited numbers of pages, only the findings of first aim are given.

Nonetheless, there are some issues that must be addressed for decision tree analyses made in this study. First of all, although CHAID performs the best accuracy on average values, it does not mean that CHAID is the best at all. As can be seen, in two cases, MSDT performs the best prediction. If we consider these two findings it can be seen than, multi-splitting decision tree approaches stand more suitable than binary-splitting techniques for our problem. Additionally, as stated before, multi-way splitting decision tree models are clearly more explainable.

By this study, it's shown that food demand prediction can be done by decision tree methods in public food courts which have definite menus and uncertain numbers of customers for each day. On the other hand, demands for menu combinations that are not presented yet can even be predicted by this approach. As a consequence, decision tree models and decision support systems developed with decision trees have significant potential for decision makers in resource optimization, planning of supply line and financial forecasting.

As a future work, authors are planning to apply SVM and ANN techniques over the dataset and develop a decision support system based on best suited algorithm after a comparative study. Furthermore, in the future, authors are planning to build models with a dataset of three year period to achieve more accurate results.

Acknowledgement

The authors wish to thank office of Health, Culture and Sport center of Hacettepe University, for their altruistic support in collection stages of dataset.

References

1. A.S. Bozkir, E. Sezer, Usage of data mining techniques in discovering the food consumption patterns of students and employees of university, International Symposium on Engineering and Architectural Sciences of Balkan, Caucasus and Turkic Republics, 2009, 104-109.
2. H.A. Nefeslioglu, E. Sezer, C. Gokceoglu, A.S. Bozkir and T.Y. Duman, Assesment of landslide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey, Mathematical Problems in Engineering, (2010), DOI: 10.1155/2010/901095.
3. R. Abiyev, V.H. Abiyev, C. Ardil, Electricity consumption prediction model using neuro-fuzzy system, World Academy of Science, Engineering and Technology, 8 (2005), 128-131.
4. M.A. Yurdusev, M. Firat, M. E. Turan, Generalized regression neural networks for municipal water consumption prediction, Journal of Statistical Computation and Simulation, 80 (2010), 477-478.
5. A. Altunkaynak, M. Özger, M. Çakmakcı, Water consumption prediction of Istanbul city bu using fuzzy logic approach, Water Resources Managament, 19 (2005), 641-654.
6. L.L. Bhattarcharyya, S. Sclove, S.L. Chen, W.J. Lattyak, Data mining on time series: an illustration using fast-food restraunt franchise data, Computatiaonal Statistics & Data Analysis, 37 (2001), 455-476.
7. H. Tang, L.N. Xing, A Web-Based Data Mining System for Forest Resource Planning System, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 4(2008), 399-403.
8. M. Berry and G. Linoff, Mastering data mining: The art and science of customer relationship management, John Wiley & Sons, 2000.
9. Z. Tang and J. MacLennan, Data Mining with Sql Server 2005, JohnWiley & Sons, 2005.
10. C.F. Chien and L.F. Chen, Data mining to improve personnel selection and enhance human capital:a case study in high-technology industry, Expert Systems with Applications, 34 (2008), 280–290.
11. K-Y. Chen and C-H, Wang, Support vector regression with genetic algorithms in forecasting tourism demand, Tourism Management, 28 (2005), 215-226.
12. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth & Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.