한국경영과학회/대한산업공학회 '97 춘계공동학술대회, 1997년 4월 25일-26일, 포항공과대학교

Integrating CBR and ID3 Data Mining Tools for Predicting Supermarket Sales

Heeseok Lee, Steven H. Kim and Jaegyong Chang

MIS, Graduate School of Management

Korea Advanced Institute of Science and Technology, Seoul

Abstract

Business organizations generate and collect a large amount of data in their daily operations. However, despite this wealth of data, many companies have failed to fully capitalize on its value because information implicit in the data is not easy to discern. In this paper, sales in the supermarket is predicted by integrating two data mining techniques such as the CBR (Case Based Reasoning) and the ID3 induction method. POS (Point of Sale) data from a real-life large retailer are analyzed. The data has chaotic and noisy time series patterns. The results of the prediction imply that the integrated method provides better prediction performance and is easy to use.

1. Introduction

The amount of information in the world is doubling approximately every 20 months, and it is becoming impossible to effectively manage that data using traditional database systems. Business organizations generate and collect large amounts of data which they use in daily operations. The data necessary for each operation is captured and maintained by the corresponding department. Yet despite this wealth of data, many companies have been unable to fully capitalize on its value because information implicit in the data is not easy to discern. However, to compete effectively today and take advantages of high-return opportunities in a timely fashion, decision-makers must be able to identify and utilize information hidden in the collected data.

Data mining is a set of analysis techniques used to uncover these nuggets of value. In recent years, data mining methods have been applied to solving marketing problem successfully. Especially, the time series prediction for economic and marketing processes is a topic of increasing interest. These data mining methods outperform statistical methods for the prediction of chaotic and noisy time series, because they are able to learn the system dependencies on their own.

In this paper, the level of sales in a supermarket is predicted by the integrated method of the CBR (Case Based Reasoning) and the ID3 induction. Analyzing the POS data set results in a non-linear problem - a chaotic and noisy time series problem. A proper forecast of the sales demand reduces stock-keeping costs because it can replenish the items according to consumer preferences and develop specific and profitable marketing strategies for the decision maker.

2. Related Work

The following four cases are most closely related with our work.

A. Prediction of sales in supermarkets by a neural network

As one of the Esprit projects, the University of Vonberg Research center in Holland predicted sales volume in a supermarket by using a neural network. The center used the back-propagation and feedforward propagation of the neural network about the sales prediction in the supermarket. It adopted the sale information of 53 articles of the same product group in a supermarket DM[8-13].

B. Marketing strategies for grouping stores for a market basket analysis

IBM consulting group stored sales data of 2000 categories or sub-categories of items for each store in the chain and attempted a market basket analysis. Market basket analysis refers to the process of examining POS data to identify affinities between

products and/or services purchased by a customer. IBM consulting group employed the association rules and neural segmentation methods in order to solve this problem[7].

C. Automatic product replenishment by Neovista's Decision Series tool

Wal-Mart was resorting to a number of approximations such as grouping stores into zones and items into groups, resulting in sub-optimal predictions and leading to occasional over-stock and under-stock conditions. Recently, Neovista has been developing the prototypes for Wal-Mart, handling the massive database and analyzing it effectively by its Decision Series tool[4].

D. Inventory control expert system for large scale retailers

In KAIST, an inventory control problem was solved by two methods, a statistical method and neural network, respectively. The test result of performance showed that the neural network approach provides the better results[20].

3. Sales Prediction Method

A. Case description

The integration of CBR and ID3 induction can be used for predicting sales in the supermarket. This method can yield better results than neural networks and is easy to use due to the characteristics of CBR - the adaptability of the domain knowledge. This paper uses sales volume during 10 weeks about the new products, which is collected by the POS system in a convenience chain store. This short term prediction doesn't consider the effect of seasonality, holidays, advertising campaigns and changing prices. These variables are found to have no effect on this short term sales data. The 49 sale information of same item group is used.

For precise forecasting, it is important to obtain accurate data set and choose the proper variables. If not, the noise would be severe[15]. In our analysis, the target variable is the sales prediction value and the input variables are the previous sales prediction levels such as increase, decrease and hold. The proper number of input variables would be determined from the pre-prediction by CBR. Raw data of sales volume are transformed to one of the increase/decrease/hold values. The increase condition is converted to the value 1. The decrease condition has the value 0, while the state of hold has the value of 0.5. The method trained on 35 cases and tested 13 cases.

B. Integration of CBR and an induction method

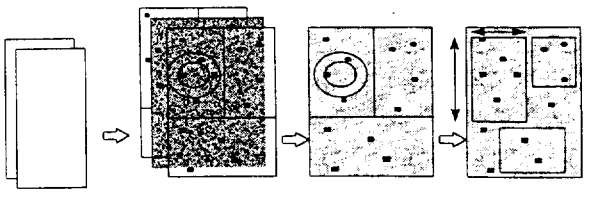
First, CBR and induction method are compared as follows.

CBR (Case Based Reasoning): The most important advantage of CBR is that it is relatively easy to acquire most of the information needed to make it work as a natural by-product of the design process. A CBR system can be effectively used even if it does not have enough domain knowledge. But, despite the fast application of CBR, it currently lacks a systematic theoretical foundation. Moreover there is not yet a consensus on good ways of resolving many of the question that occur in the design of CBR[2].

Induction method (ID3 algorithm): The ID3 algorithm determines the classification of an object by testing its values for certain properties. ID3 repeatedly partitions the training instances according to the variable with the greatest discriminatory power, using an information-theoretic measure of entropy .

CBR offers flexibility regarding the domain knowledge. Induction techniques yield general knowledge based on instances. Both CBR and the inductive method have drawbacks for smart prediction, but provide the useful predictive information for decision makers.

Therefore, integrating two methods is likely to improve the predictive quality and to enhance comprehension. In contrast, the neural network is effective but considered as black-box. The architecture in Figure 1 shows the integration of case based reasoning and the induction method. The goals of the integration are to use the tools easily, to improve the results of prediction, and to provide new knowledge for the decision maker.



<Fig 1> Integration of CBR and ID3

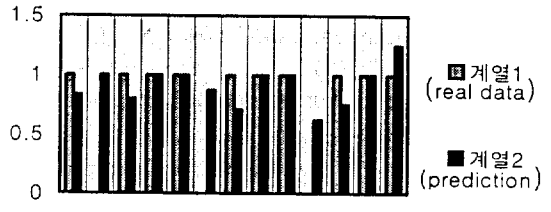
4. Results

The N-neighbor CBR method is used in the pre-prediction for further prediction. These results show which condition is optimal in terms of reduced errors. Input variables are tested at the level of 2 or 3 weeks for the selection of the optimal model. The number of training cases is 35 and that of the testing set is 14. The prediction error (MAE: Mean Average Error) is summarized below.

No. of input node \ No. of neighbors	3	4	5
2 weeks	0.61	0.29	0.27
3 weeks	0.34	**0.25**	0.36

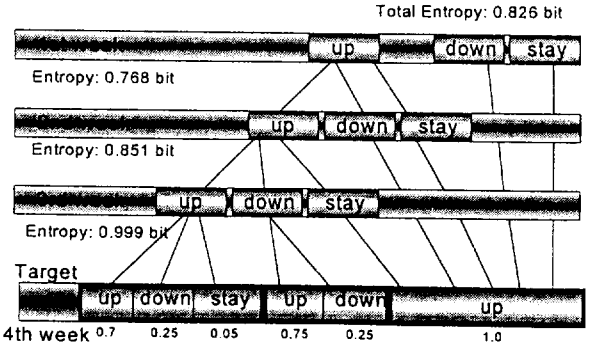
<Table 1> The pre-predictive error of N-nearest neighbor CBR

Table 1 shows that the 4 nearest neighbor model with 3 inputs is optimal for CBRs and thus 3 input node is selected for further prediction using ID3. According to this 4 nearest neighbor case, predictive result of sales is as shown in Figure 2.



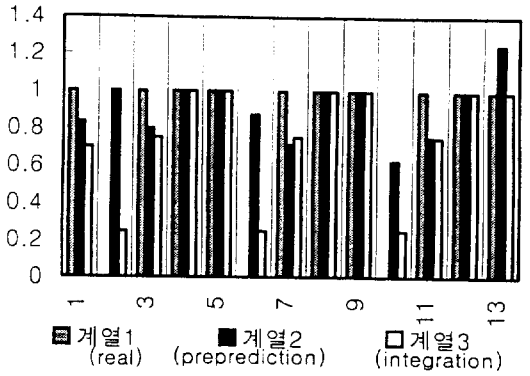
<Fig 2> Prediction Result by CBR

By the ID3 algorithm the best discriminative variable is the sales for the first week and the second best is the sales for the second week. The third best is that of the third week . The decision tree as shown in Figure 3 is generated by the ID3 procedure.



<Fig 3> Result of prediction by ID3 algorithm

Lastly, the result of the integrated method is compared and showing in Figure 4. MAE for the final result is 0.23. This error is lower than that of the optimal model of N-nearest neighbor .CBR.



<Fig 4> Result of integrated prediction

The results show that the integrated method can be applied to chaotic patterns of sales. In addition it may reduce time to predict and improve predictive quality. Furthermore, our method is fairly easy to use because CBR provides flexibility to accommodate incremental domain knowledge.

5. Conclusions

CBR is useful as a pre-prediction tool. CBR readily learn from new domain knowledge. However, it does not guarantee accurate results. ID3 is better for understanding the decision behind a prediction but fails to control the distribution of data

This paper proposes a novel approach to integrate CBR with ID3 to complement the two methods. Our method is applied to analyze supermarket data and the preliminary results are promising. We are currently in the process of sharpening our method and applying the integrated technique to larger data sets. We shall report any interesting results in the near future.

Reference

[1] 이재규, 최형림 등, 전문가시스템: 원리와 개발 (서울, 법영사, 1996)

[2] S. H. Kim and M. B. Novick "Using clustering techniques to support case reasoning", *International Journal of Computer Applications in Technology,* Vol.6, No 2 (1993), pp.57-73

[3] G. F. Luger and W. A. Stubblefield, Artificial Intelligence, (John Wiley ,1993)

[4] Decision Series White Paper (http://www.neovista.com/Software/WP_DecisionSeries.htm)

[5] Predicting Sales of Articles in Supermarkets, *Stimulation Initiative for European Neural Applications Esprit Project 9811: Case Studies of Successful Applications* (http://septimius.mbfys.kun.nl:80/snn/siena/cases/supermarkets.html)

[6] Modelling Market Dynamics in Food-, Durables- and Financial Markets, *SIENA: Stimulation Initiative for European Neural Applications Esprit Project 9811* (http://septimius.mbfys.kun.nl:80/snn/siena/cases/brandmarc.html)

[7] Reality Check for Data Mining (http://www.almaden.ibm.com/stss/papers/reality/)

[8] Short Term Prediction of Sales in Supermarkets, *Proceedings ICNN'95, Perth, Western Australia 27 Nov - 1 Dec '95, Vol 2 of 6, pp. 1028-1031*

[9] Parallel Back-Propagation for Sales Prediction on Transputer Systems, *Procs. World Transputer Congress '95, 4 - 6 September 1995, Harrogate, UK, IOS Press 1995, pp.318-331*

[10] A Neural Network Approach for Predicting the Sale of Articles in Supermarkets, *Procs. EUFIT '95, 3rd European Congress on Intelligent Techniques and Soft Computing, 28-31 August 1995, Aachen, Germany, Vol 1 of 3*

[11] Time Series Prediction by Neural Networks , *Power Xplorer User Report, Parsytec GmbH, May 1995, pp.33-36*

[12] Parallel Backpropagation for the Prediction of Time Series , *First European PVM Users' Group Meeting, Rome, Italy, Oct. 9-11, 1994*

[13] Parallele Backpropagatio Parallele Datenverarbeitung aktuell: *TAT'94, 26./27. Sept. 1994, IOS Press, 1994, pp.419-427*

[14] Introduction to Data Mining (http://godard.oec.uni-osnabrueck.de/fachgeb/winf2/fachinfo/datam01.html)

[15] Data Collection and Sparse Data Issues (http://www.trajecta.com/caseedu1.html)

[16] Klaus-Dieter Althoff, S.Wess and Ralph Traphoner, INRECA:- a Seamless Integration of Induction and Case based Reasoning for Decision Support Tasks. *Induction and Reasoning from Cases(ESPRIT P6322), University of Kaiserslautern*, 1995 (http://wwwagr.informatik.uni-kl.de/~lsa/CBR/INRECAproject.html)

[17] J. J. Daniels, Retrieval of Passages for Information Reduction. *NSF Grant No. EEC-9209623 July 19,1996* (http://www.aic.nrl.navy.mil/~breslow/cbr/research-summaries.html#)

[18] Reasoning with Case Bases (http://www.haley.com/cbrcomp.html)

[19] M.M Richardson and, J.R. Warren., Expert System Construction by Reasoning from Cases . University of South Australia, 1996 (http://www.cis.unisa.edu.au/projects/escrccs.html)

[20] Kwang Yon Lee, Inventory Control Expert System for Large Scale Retailers. *KAIST Masters Thesis, 1995*

[21] M. Manage, Building a Corporate Memory for Decision-Making (http://www2.cordis.lu/esprit/src/results/pages/infoind/infind8.html)