Integration of Heterogeneous Models with Knowledge Consolidation

Jinhwa Kim[a], Jae Kwon Bae[b]

[a, b] *School of Business, Sogang University*
#1 Sinsu-dong, Mapo-gu, Seoul 121-742, KOREA
Tel: +82-2-705-8860, Fax: +82-2-705-8519, E-mail: {jinhwakim, baejaekwon}@sogang.ac.kr

Abstract

For better predictions and classifications in customer recommendation, this study proposes an integrative model that efficiently combines the currently-in-use statistical and artificial intelligence models. In particular, by integrating the models such as Association Rule, Connection Frequency Matrix, and Rule Induction, this study suggests an integrative prediction model.

Keywords:

Customer Recommendation, Artificial Intelligence, Rule Induction, Integrative Prediction Model

Introduction

This study suggests an integrative model to produce synergy effects from combinations of heterogeneous models. It uses knowledge in the form of rules as a meta model to integrate different models such as Association Rules, Connection Frequency Matrix, and Rule Induction System. The models in these methods are converted into rules and these rules are integrated into a single meta model. The data set from a convenience store *G* in S. Korea is collected for the tests. The study compares its performance with that of other single type models for validation.

The remainder of this paper is organized as follows. Section 2 introduces literature review. Section 3 explains the processes of data collections and variables selections. Section 4 shows three heterogeneous prediction models and a suggested model by integrating these three models. Section 5 summarizes and analyzes empirical results. Section 6 discusses the conclusions and future research issues.

Prediction of Purchasing Intention and Recommendation Systems

Customer buys some products with his purchasing intention arose from solvency. Therefore, prediction of customer's purchasing intention is a very interesting issue in marketing sector. If companies have abilities of grasping about customer's purchasing behavior and personalized product recommendation system using his information, they can utilize various chances of business [8][16].

Recommendation system provides advice to customer about items they might wish to purchase or examine. Recommendation made by such systems can help users navigate through large information spaces of product descriptions, news articles or other items. As on-line information and e-commerce burgeon, recommender systems are an increasingly important tool [7]. Personalized product recommendation systems recommend product and service based on customer's private information or purchasing behavior in store. Burke [6] describes three different types of recommendation approaches: collaborative-filtering, content-based and knowledge-based.

The most well known type of recommender system is the collaborative-filtering type. These systems aggregate data about customer's purchasing habits or preferences, and make recommendation to other users based in similarity in overall purchasing patterns.

Content-based recommender systems are classifier systems derived from machine learning research. For example, the NewsDude news filtering system is a recommender system that suggests news stories the user might like to read [4]. These systems use supervised machine learning to induce a classifier that can discriminate between items likely to be of interest to the user and those likely to be uninteresting.

A third type of recommender system is one that uses knowledge about users and products to pursue a knowledge-based approach to generating a recommendation, reasoning about what product meet the user's requirements. The PersonalLogic recommender system offers a dialog that effectively walks the user down a discrimination tree of product features [6][15].

Researches about the latest personalized product recommendation system using collaborative filtering technique are progressed by application of various data mining algorithms. The representative research is Association rule. And there is also Neural network which is operated for product recommendation using prediction of customer's purchasing intention [5].

Data Collection

The data set for the tests is collected from a convenience store *G*, which is the number one in its brand in S. Korea.

This data set contains sales information on customer transactions from September 1, 2005 to December 7, 2005. About 1,000 transactions are selected for a specific item. The data has information on Date_Purchased, Time_Purchased, POS, Employee Number, Receipt Number, Item_Name, Quantity, Purchase_Cost, and Description. Using this data set, it suggests an integrated method predicting whether a customer buys or not buys a specific product for target marketing strategy.

The items in the data set are classified into 21 categories based on their characteristics as in Table 1.

Table 1 - Definition of variables

Category	Definition	Category	Definition
1	Processed food	12	Household items
2	Health beverage	13	Distilled liquor
3	Sweets	14	Newspaper
4	Rice rolled in dried laver	15	Yogurt
5	Frozen food	16	Milk
6	Tobacco	17	Juice
7	Instant noodle	18	Chocolate
8	Beer	19	Candy
9	Ice cream	20	Coffee
10	Bread	21	Carbonated beverage
11	Spring water		

The data set is divided into two subsets, a training data set and a test (holdout) data set. The training data set is used to train the prediction models. The test data set is used to test the performance of the suggested model. For each test data set, a training subset and test (holdout) subset, consisting of 80% and 20% of the data, respectively, are randomly selected. The tests have been replicated ten times to reduce the random impact in composition of data set [18].

Suggested Models

Association rule

A typical association rule has an implication of the form A → B where A is an item set and B is an item set that contains only a single atomic condition. The support of an association rule is the percentage of records containing item sets A and B together. The confidence of a rule is the percentage of records containing item set A that also contain item set B. Support represents the usefulness of discovered rules and the confidence represents certainty of the detected association rule [17].

Association rule mining technique finds all collections of items in a database whose confidence and support meet or exceed pre-specified threshold values. Apriori algorithm is one of the prevalent techniques used to find association rules [1]. Apriori operates in two phases. In the first phase, all large item sets are generated. This phase utilizes the downward closure property of support. The second phase of the algorithm generates rules from the set of all large item sets.

Association rule technique is implemented with SAS E-Miner 4.0. with support of 0.5%, and confidence of 10%. There are 11 rules induced from the test with the first data set as in Table 2.

Table 2 - Generalized rules from data set 1

Support (%)	Confidence (%)	Generalized Rule	Explanations
5.5	82	18 → 16	Chocolate → Milk
3.5	86	1 → 16	Processed food → Milk
3.5	57	6 → 19	Tobacco → Candy
2	50	2 and 10 → 15	Health beverage and Bread → Yogurt
1.5	67	3 and 10 → 8	Sweets and Bread → Beer
1.5	67	14 → 6	Newspaper → Tobacco
1	100	16 and 19 → 18	Milk and Candy → Chocolate
1	50	9 → 12	Ice cream → Household items
1	50	3 and 5 → 7	Sweets and Frozen food → Instant noodle
1	50	2 and 5 → 20	Health beverage and Frozen food → Coffee
1	50	9 → 3	Ice cream → Sweets

CFM(Connection Frequency Matrix)

CFM(Connection Frequency Matrix) can be used to find correlations among products from the analysis of the scored frequencies in the frequency matrix. These correlations among products can be represented in the form of rules. For example, if a transaction *N* has 3(Sweets) and 7(Instant noodle) in a basket, one point is added to the cross section of row 3 and column 7 in the frequency matrix as in Figure 1. If a transaction has 2(Health beverage), 1(Processed food), and 6(Tobacco) together, the products are sorted by their number as {1,2,6}, and the whole set is divided into three pairs of {1,2},{1,6},{2,6}. The point of one is added for these three pairs as in Figure 1. This frequency matrix shows connectivity or correlations among products with numbers. The numbers in the matrix show that the bigger the number between products, the greater the chance of co-purchase.

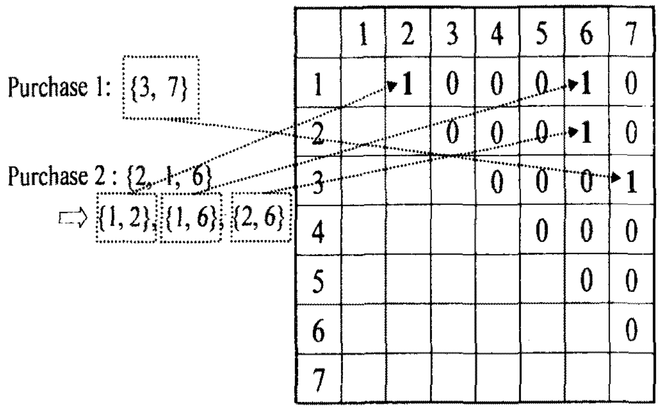

Figure 1 - Learning/training algorithm of the CFM

The test is performed on milk as it is the most frequently purchased item the store. It predicts whether a customer will buy or not buy milk. From original data set with 1000 transaction records, 800 records are used for training data set and 200 records are reserved for the test data set. The training data set has 400 records which contain milk, and the other 400 records which do not contain milk. CFM technique is implemented with Visual Basic .Net of Microsoft.

Rule Induction

Rule induction refers to the rules derived from the decision tree technique in data mining. The data set is separated into many partitions in a way to increase the purity, which is the degree to which the dependent variable belongs to a certain class. The rules that are applied for splitting the data are called the inducted rules. Rule induction is a non-parametric method and suitable for figuring out interaction effect or non-linearity. In many cases, decision tree is used for the sake of interpretation of the analysis results acquired by the neural networks.

Quinlan's ID3 has been widely reported in the literature and is amongst the more commonly used induction algorithms available in commercial implementations. We use the C5.0 programs that incorporate enhancements to the basic algorithm to consider continuous variables and also includes pruning.

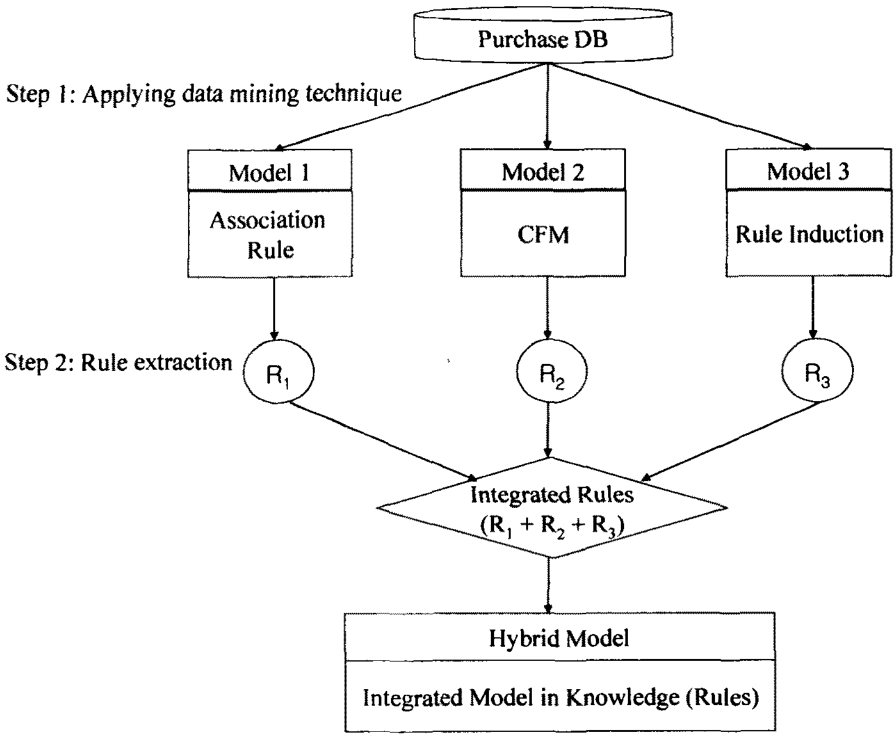

Figure 2 - The processes of building an integrated model from 3 models

An Integrated Model as a Suggested Method

There are two steps in building the integrated model. Step 1 produces models from the transaction data set with Association Rule, CFM, and Rule Induction technique. Step 2 converts each of these three models into a unified form of rules. Rules from each of these heterogeneous models are merged into one integrated model. The rules (R_1, R_2, R_3) in step 2 are merged ($R_1+R_2+R_3$) into an integrated model as in Figure 2.

Experiments Results

Table 3 and Figure 3 compare the prediction performances of Association Rule, CFM, Rule Induction, and the integrated model from 10 fold cross validations.

Among these methods, the integrated model has the highest level of average accuracy (75.35%) with given data sets, followed by Rule Induction (67.6%), and Association Rule (64.0%) next.

Clearly, the integrated model shows, on average, superior prediction accuracy compared to other single models. Also, the performance of the suggested model is better than other models in all ten tests.

Table 3 – Performance Comparison of 4 models

	Association Rule	CFM	Rule Induction	Integrated Model
Set 1	64.0	52.5	68.0	73.5
Set 2	62.5	51.5	69.5	77.5
Set 3	68.5	51.0	73.5	79.5
Set 4	66.5	52.0	73.0	79.0
Set 5	61.5	51.0	64.5	70.0
Set 6	67.0	50.5	70.0	82.5
Set 7	62.0	52.0	68.0	70.0
Set 8	64.5	52.5	64.0	77.0
Set 9	59.0	52.0	64.5	74.0
Set 10	64.5	52.0	61.0	70.5
Mean	**64.0**	**51.7**	**67.6**	**75.35**
Std	2.85	0.67	4.07	4.40

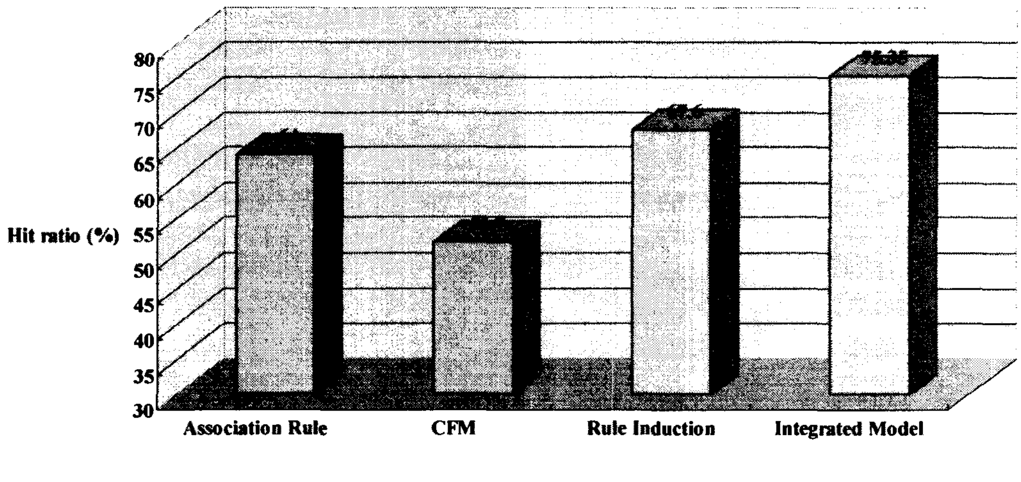

Figure 3 - Performance of the integrated model

Table 4 shows the results of McNemar tests to evaluate the classification performance of the suggested model. As shown in the Table 4, the integrated model performs significantly better than every single model proposed for this study at the give type one error of 5% or 10%.

Table 4 – McNemar test

	Association Rule	CFM	Rule Induction
Integrated Model	8.481 (0.004)**	44.180 (0.000)**	2.813 (0.094)*
Association		24.038	0.291[1]

Rule	(0.000)**	(0.590)[2]
CFM		10.223 (0.001)**

1. Chi-square value / 2. P-value
*significant at 10% / **significant at 5%

Conclusions

This study suggests an integrative approach combining three different prediction models. It uses knowledge as a uniform representation of these three models. The rules are derived from each model, and these rules are merged together for one integrative model. The results from the experiments show that the performance of the suggested method is superior to that of other methods such as Association Rule, CFM, and Rule Induction. Future studies will try to include more models such as regression models and artificial neural networks into the integrated model.

References

[1] Agrawal, R., Imielinski, T., and Swami, A. (1993). "Mining association rules between sets of items in large databases," *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207-216.

[2] Altman, E. I., Marco, G. and Varetto, F. (1994). "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," *Journal of Banking and Finance*, Vol. 18, No. 3, pp. 505-529.

[3] Berry, M., Linoff, G. (1997). *Data Mining Techniques*. New York: Wiley.

[4] Billsus, D. and Pazzani, M. (1999). "A Hybrid User Model for News Story Classification," *In Proceedings of the Seventh International Conference on User Modeling*.

[5] Brijs, T., Swinnen, G., Vanhoof, K. and Wets, G. (1999). "Using Association Rules for Product Assortment Decisions: A Case Study," *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 254–260.

[6] Burke, R. (1999). "The Wasabi Personal Shopper: A Case-Based Recommender System," *Proceedings of the 11th Conference on Innovative Applications of Artificial Intelligence*, pp. 844-849.

[7] Burke, R. (2000). "Knowledge-based Recommender Systems," *Encyclopedia of Library and Information Systems*, Vol. 69, Supplement 32.

[8] Chiu, C. (2002) "A case-based customer classification approach for direct marketing," *Expert Systems with Applications*, Vol. 22, No. 2, pp. 163–168.

[9] Desai, V. S., Conway, D. G., Crook, J. N., and Overstreet, G.A. (1997). "Credit scoring models in the credit union environment using neural networks and genetic algorithms," *IMA Journal of Mathematics Applied in Business and Industry*, Vol. 8, pp. 323–346.

[10] Lee, K.C., Kim, M.J., Kim, H. (1994). "An Inductive Learning – Assisted Neural Network Approach to Bankruptcy Prediction: Comparison with MDA, Inductive Learning, and Neural Network Models," *Journal of Management Research*, Vol. 23, No. 3, pp. 109–114.

[11] Lee, G.C., Han, I.G., Kim, M.J. (1996). "A Study on the Credit Evaluation Model Integrating Statistical Model and Artificial Intelligence Model," *Journal of Management Science*, Vol. 21, No. 1, pp. 81–100.

[12] Mangasarian, O.L. (1965). "Linear and nonlinear separation of patterns by linear programming," *Operations Research*, Vol. 13, No. 3, pp. 444–452.

[13] Ohlson, J.A. (1980). "Financial ratios and probabilistic prediction of bankruptcy," *Journal of Accounting Research*, Vol. 18, No. 1, pp. 109–131.

[14] Peel, M.J., Peel, D.A., and Pope, P.F. (1986). "Predicting corporate failure-some results for the UK corporate sector," *Omega International Journal of Management Science*, Vol. 14, No. 1, pp. 5–12.

[15] Ricci, F. and Del Missier, F. (2003). "Personalized Product Recommendation through Interactive Query Management and Case-Based Reasoning," *In Proceedings of CHI'03 Workshop on Designing Personalized User Experiences for eCommerce.*

[16] Schafer, J., Konstan, J., and Riedl, J. (2001). "E-commerce Recommendation Applications," *Data Mining and Knowledge Discovery*, No. 5, No. 1, pp. 115–153.

[17] Song, H.S., Kim, S.H., Kim, J.K. (2001). "A Methodology for Detecting the Change of Customer Behavior based on Association Rule Mining," *Proceedings of the PACIS 2001*, pp. 871-885.

[18] Weiss, S., Kulikowski, C. (1991). *Computer Systems That Learn*. Morgan Kaufmann Publishers, Inc.

[19] West, D. (2000). "Neural Network Credit Scoring Models," *Computers and Operations Research*, Vol. 27, pp. 1131–1152.

[20] Wiginton, J.C. (1980). "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior," *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3, pp. 757–770.

[21] Zmijewski, M.E. (1984). "Methodological Issues Related to the Estimation of Financial Distress Prediction Models," *Journal of Accounting Research*, Vol. 22, No. 1, pp. 59–82.