A study of extended RFM analysis based on PLSA model for Purchase History Data

Qian Zhang
Graduate School of Creative Science and Engineering
Waseda University, Tokyo, Japan
Tel: (+81) 03-5286-3290, Email: cho-sei@asagi.waseda.jp

Haruka Yamashita
School of Creative Science and Engineering
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: h.yamashita@aoni.waseda.jp

Kenta Mikawa
Department of Information Science
Shonan Institute of Technology, Kanagawa, Japan
Tel: (+81) 466-30-0212, Email: mikawa@info.shonan-it.ac.jp

Masayuki Goto
School of Creative Science and Engineering
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: masagoto@waseda.jp

Abstract. On the highly developed information society, it has become more important to analyze customers' purchase history to know customers preferences. In the field of marketing, the RFM (recency, frequency, and monetary) analysis has been widely used, and in the field of machine learning, probabilistic latent semantic analysis (PLSA) is well known as a soft clustering method. Recently, the RFM analysis based on PLSA model was proposed and it enables us to cluster customers into latent classes, to show the expectation value of recency, frequency and monetary, and to calculate the assignment probabilities of each customer to each latent class. This paper adds a new factor "when the customer purchased first" to the three factors of RFM analysis based on PLSA model. We call the method extended RFM analysis based on PLSA. Moreover, we visualize the result used extended RFM analysis based on PLSA model using self-organizing map (SOM) which visualizes the structure of the data and also divides customers into several groups. For demonstrating an example of an analysis based on our proposal, we analyzed customers' purchase history data using the RFM analysis based on PLSA, the extended RFM analysis based on PLSA, and the combined method with SOM.

Keywords: segmentation, RFM analysis, latent class, SOM

1. INTRODUCTION

On the highly developed information society, it has become more important to analyze customers' purchase history to know customers preferences. Market segmentation by using the customers' purchase history is especially useful in the marketing field to devise various strategies to improve a business' performance (Beane & Ennis, 1987). As one of the methods for analyzing purchase data, the RFM analysis (Bult & Wansbeek, 1995; Birant, 2011) is widely known. It express customers' preference and typically used to segment customers into several groups by using three variables: how long it has been since their last purchase (Recency), how many times they purchased (Frequency), and the quantity they spent (Monetary) (Tsai & Chiu, 2004; Khajvand, 2011; Wei et al., 2012). On the other hand, the authors proposed a new latent class model for the RFM analysis to represent customer purchasing behaviors based on its three variables (Zhang et al., 2015). It enables us to cluster customers into latent classes, and analyze the characteristics of the latent classes. However, the information about the date of the first purchase by each customer has not been incorporated in the model, but is important for a retail store. For example, monitoring yearly purchase data of customers, there are customers both of their R is within a week, F is 10 times, M is 10,000 yen and one's data of the first purchase is with a month, and other is earlier than half of year, their clusters should be different (i.e., one is in new customers' cluster, and another is in a good customers' cluster). Hence, the information about the first purchasing action by customers enables to analyze customer better for marketing.

Moreover, since the output of RFM analysis based on latent class model cannot be visualized in the 2-dimensional space, it is also desired that the analysis result based on the latent class model is visualized in a 2-dimensional space to represent the result transparently and understandably.

In this paper, we add a new factor "when the customer purchased first" to the three factors of the RFM analysis based on PLSA model. We call this method the extended RFM analysis based on PLSA. Moreover, we visualize the result of the extended RFM analysis based on PLSA model by applying the self-organizing map (SOM) which enables the visualization of the feature of the data and also division of customers into several groups (Hanafizadeh & Mirzazadeh, 2011). For demonstrating our extended model, we analyze a real data from a major Japanese retail company using the RFM model based on PLSA, the extended RFM model based on PLSA, and visualize the results of the analysis by SOM. This data is i_code data from ID's COOPERATION and provided by the 2015 Data Analyzing Competition, held by the Joint Association Study Group of Management Science in Japan.

2. THE RFM ANALYSIS BASED ON PLSA MODEL

2.1 Probabilistic Latent Semantic Analysis (PLSA)

The probabilistic latent semantic analysis (PLSA) is widely used for soft clustering problems (Hofmann & Puzicha, 1999; Hofmann, 1999). It is a technique for one of the topic models, and it was initially used for text-based applications, such as information retrieval or text clustering. This model is one of probabilistic latent class models and several different unobserved classes can be assumed behind observed variables. When applied to the customer segmentation it assumes latent classes between users who have similar preferences, and product items that have a similar purchase tendency. This model additionally assumes that the users and the product items belong to each latent class stochastically; that is, it allows that they belong to several different latent classes. The diversity of the user preferences and the tendency of product items are represented based on this assumption. Here, let u_r $(r = 1, \ldots, m)$ be user, a_j $(j = 1, \ldots, n)$ be the product item, and z_k $(k = 1, \ldots, K)$ be the latent class. The graphical model of PLSA is described in Figure 1.

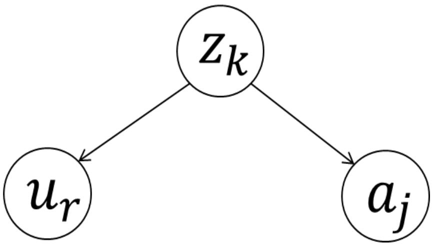

Figure 1: Graphical model of PLSA.

The co-occurrence event of the user u_r and the product item a_j can be modeled by the probabilities $P(z_k)$ and the conditional probabilities $P(u_r|z_k)$ and $P(a_j|z_k)$ in the PLSA. The probabilistic model is formulated by the following equation:

$$\mathrm{P}(u_r, a_j) = \sum_{k=1}^{K} P(z_k)\mathrm{P}(u_r|z_k)\,\mathrm{P}(a_j|z_k) \ , \qquad (1)$$

where $P(z_k)$ satisfies $\sum_{k=1}^{K} \mathrm{P}(z_k) = 1$.

2.2 The PLSA Model for RFM Analysis

Recently the modified PLSA model using the feature variables of R, F, M variables simultaneously for customer segmentation has been proposed (Zhang et al., 2015). It assumes latent classes between the three variables of customers, who have similar preferences, and enables to show the expectation value of recency, frequency and monetary, and to calculate the assignment probabilities of each customer to each latent class.

Here, let the R, F, M variables are denoted by $x_{ni}(i \in \{1, \ldots, L\}, n = \{1,2,3\})$, where $n = 1$ is a representation of the R variable, $n = 2$ is a representation of the F variable, and $n = 3$ is a representation of the M variable respectively, and $z_k(k = 1, \ldots, K)$ be the latent classes. The purchase behavior of the customer i is then denoted by a vector (x_{1i}, x_{2i}, x_{3i}) . The proposed probabilistic model is formulated by Equation (2):

$$\mathrm{P}(x_{1i}, x_{2i}, x_{3i}) = \sum_{k=1}^{K} P(z_k)P(x_{1i}|z_k)\,P(x_{2i}|z_k)(x_{3i}|z_k) \ . \tag{2}$$

Here, $P(x_{ni}|z_k)$ are the conditional probabilities of the R, F, and M variables, conditioned by a latent class z_k respectively. This model assumes a normal distribution of the conditional probabilities $P(x_{ni}|z_k)$. Its graphical model is described as follows.

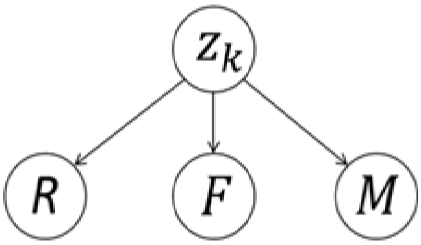

Figure 2: RFM analysis based on PLSA model.

Denoting its average as μ_{nk} and its variance as σ_{nk}^2, and the probabilities $P(x_{ni}|z_k)$ can be represented by the following equation:

$$P(x_{ni}|z_k) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{(x_{ni}-\mu_{nk})^2}{2\sigma_{nk}^2}\right] \ . \tag{3}$$

These parameters μ_{nk}, σ_{nk}^2, $P(z_k)$ are estimated by the EM algorithm.

3. EXTENDED RFM ANALYSIS BASED ON PLSA MODEL

The RFM analysis based on PLSA model has focused on the recency, the frequency, and the monetary. However, according to the factor "when the customer purchased first", the customers' interpretation is greatly changed. If there are customers who have the same values of R, F, and M, and the first purchase date is greatly different; the customers should belong to different groups. For example, monitoring yearly purchase data of customers, there may be customers whose R is within a week, F is 10 times, M is 10,000 yen. If the date of the first purchase by a customer is with a month, but that of another customer is earlier than half of year, then their clusters should be different (i.e., one should be in new customers' cluster, and another should be in a good customers' cluster). In this study, we add the new factor "when the customer purchased first" to the three factors of the RFM analysis based on PLSA model.

3.1 Formulation of the extended model

Here, let the new factor of "the date of first purchase by the customer" (FIRST) be denoted by x_{4i}. The conditional probability of the FIRST variables is denoted by $P(x_{4i}|z_k)$, and the extended probabilistic model is formulated by Equation (4):

$$P(x_{1i}, x_{2i}, x_{3i}, x_{4i}) = \sum_{k=1}^{K} P(z_k)P(x_{1i}|z_k)\,P(x_{2i}|z_k)(x_{3i}|z_k)(x_{4i}|z_k) \quad . \tag{4}$$

The conditional probabilities $P(x_{4i}|z_k)$ are estimated in the same way to the $P(x_{ni}|z_k)$ of Equation (3) by assuming a normal distribution.

3.2 The Method of Parameter Estimation

The parameters presented in the subsection 3.1 can be estimated from the purchase history data to maximize the log-likelihood. The log-likelihood function LL is defined by the following equation:

$$LL = \sum_{i=1}^{N} logP(x_{1i}, x_{2i}, x_{3i}, x_{4i}) \ . \tag{5}$$

Note that $P(x_{1i}, x_{2i}, x_{3i}, x_{4i})$, given by Equation (4) includes the latent variables z_k, which cannot be observed. Since the estimator of these model parameters, the latent variable z_k cannot be calculated analytically, it is necessary to employ an iterative procedure, such as the EM algorithm. The EM algorithm is the method that estimates the parameter based on the maximum likelihood principle, by using an iterative procedure with only the observed data. This algorithm contains two steps, which are the expectation step (E-step) and the maximizing step (M-step). The E-step calculates the conditional expectation from the observed data. The M-step maximizes the conditional expectation of the log-likelihood, which is calculated in the previous E-step. By iterating these two steps, the log-likelihood finally converges to a local maximum and the estimated parameters are then produced. Each step of this algorithm for the proposed model is formulated by the following equations:

[The E-step]

$$P(z_k|x_{1i}, x_{2i}, x_{3i}, x_{4i}) = \frac{P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)P(x_{4i}|z_k)}{\sum_{k=1}^{K} P(z_k)P(x_{1i}|z_k)\,P(x_{2i}|z_k)P(x_{3i}|z_k)P(x_{4i}|z_k)} \ . \tag{6}$$

The probabilities $P(z_k|x_{1i}, x_{2i}, x_{3i}, x_{4i})$ are updated after estimating each parameter in the M-step.

[The M-step]

$$P(z_k) = \frac{1}{N} \sum_{i=1}^{N} P(z_k | x_{1i}, x_{2i}, x_{3i}, x_{4i}) \ , \qquad (7)$$

$$\mu_{nk} = \frac{\sum_{i=1}^{N} x_{ni} P(z_k | x_{1i}, x_{2i}, x_{3i}, x_{4i})}{\sum_{i=1}^{N} P(z_k | x_{1i}, x_{2i}, x_{3i}, x_{4i})} \ , \qquad (8)$$

$$\sigma_{nk}^2 = \frac{\sum_{i=1}^{N} (x_{ni} - \mu_{nk})^2 P(z_k | x_{1i}, x_{2i}, x_{3i}, x_{4i})}{\sum_{i=1}^{N} P(z_k | x_{1i}, x_{2i}, x_{3i}, x_{4i})} \ . \qquad (9)$$

The EM algorithm is iterated when the log-likelihood function of the equation (5) convergence.

4. ANALYSIS OF PURCHASE HISTORY DATA BASED ON EXTENDED MODEL

For the presentation of the applicability of our model for the real-world marketing data, in this section, we analyze the real purchase history data of a major Japanese retail company by applying the proposed method and conventional method (Zhang, et al, 2015) in order to compare their performance. The data for this demonstration was provided by the 2015 Data Analyzing Competition, held by the Joint Association Study Group of Management Science in Japan.

4.1 Data and Analysis Settings

The proposed method is applied to real purchase history data to verify the effectiveness of our proposal. We used purchase history data stored from January 1, 2014, to December 31, 2014 stored in a store from a major Japanese retail company. The number of customers is $I = 111{,}753$. In addition, the number of latent classes is set to $K = 20$, because of the empirical method in advance. To evaluate each latent class, we use the averages of each 4 variable of each latent class. The detailed results are described in the subsection 4.2.

4.2 Result and Discussion

The results of the proposed model and the RFM analysis based on PLSA model are shown in Table1 and Table2. Each column of Tables 1, 2 presents "how long it has been since their first purchase", "how long it has been since their last purchase" (the smaller number is, the more it close to the present day), "how many times they purchased during one year", "how much they spent for their purchase during one year", and "customers' rate of belonging in the latent class", respectively.

Table 1: Result of the proposed model.

	FIRST	R	F	M	percent
latent class1	183.7	183.7	1.0	1267.0	33.64%
latent class2	230.1	135.4	2.0	2584.9	14.57%
latent class3	254.5	107.6	3.0	3969.8	8.47%
latent class4	215.5	44.7	4.0	5927.0	2.91%
latent class5	323.2	138.2	4.0	4854.8	2.89%
latent class6	282.6	80.1	5.0	6909.6	4.21%
latent class7	280.3	52.8	6.5	6511.8	2.47%
latent class8	339.4	148.7	7.6	8246.3	2.40%
latent class9	226.3	13.9	8.8	10423.7	3.09%
latent class10	349.0	36.3	10.1	11674.2	3.28%
latent class11	280.1	77.1	12.5	18488.0	2.33%
latent class12	340.7	16.0	16.0	21301.7	3.91%
latent class13	357.3	136.7	17.8	22693.6	2.17%
latent class14	222.2	8.2	21.3	28605.4	1.91%
latent class15	354.8	11.2	31.9	44271.8	4.29%
latent class16	360.9	102.4	52.2	71734.4	0.90%
latent class17	278.2	4.8	57.2	82986.8	1.06%
latent class18	359.7	5.6	74.2	108172.0	3.31%
latent class19	313.2	51.9	96.6	159470.0	0.62%
latent class20	362.5	2.4	182.1	314337.0	1.55%

Table 2: Result of the conventional model.

	R	F	M	percent
latent class1	183.8	1.0	1261.5	33.62%
latent class2	135.4	2.0	2583.5	14.56%
latent class3	107.6	3.0	3968.8	8.47%
latent class4	91.4	4.0	5370.1	5.79%
latent class5	80.1	5.0	6909.7	4.21%
latent class6	73.0	6.0	8371.8	3.10%
latent class7	102.9	7.3	6658.9	1.53%
latent class8	9.7	8.4	8766.3	1.79%
latent class9	68.3	9.2	10771.2	2.32%
latent class10	24.9	11.2	12903.6	3.05%
latent class11	203.3	12.3	14676.9	1.54%
latent class12	8.2	15.6	19163.1	3.70%
latent class13	53.2	16.0	22780.8	2.50%
latent class14	19.6	25.5	36713.3	2.78%
latent class15	4.6	31.6	40820.5	2.89%
latent class16	128.8	36.3	48021.0	1.21%
latent class17	8.1	55.3	80366.0	3.21%
latent class18	73.0	104.2	163793.0	0.46%
latent class19	3.8	107.5	165763.0	2.37%
latent class20	1.8	219.7	397618.0	0.89%

Tables 1 and 2 are the estimated averages of each variables conditioned by a latent class evaluated based on the extended model and the previous model (Zhang, et al., 2015) ordered by ascending frequency. For example, the latent classes 1 and 2 of the two tables are the groups of non-prime customers because the averages of their R variables are larger than the values of other classes, and the averages of their F and M variables are smaller. Adversely, the latent class 20 of Tables 1 and 2 is the best customer because the averages of their R variables are smallest, and the averages of their F and M variables are largest in the all classes.

In addition, focus on the latent classes 4 and 5 of the Table1. Both of the averages of the F variable are 4.0; however, other averages are different; the average of FIRST and R variables of latent class 4 are smaller, and the average of the M variable is larger than that of the latent class 5. Hence, the latent class 4 is a group of usual customers, and the latent class 5 is a group of estranged customers. On the other hand, on the latent class 4 of the conventional model, the averages of the three variables are the middle valued between the latent classes 4 and 5 of proposed one. These results suggest that the proposed model is more suitable to analyze customers.

From the above discussions, we clarified that it is possible to segment estranged customers by adding the new factor “when the customer purchase first” which is important for a retail store as a factor.

5. VISUALIZATION OF EXTENDED MODEL

The Self-Organizing Map (SOM) is a popular unsupervised neural network methodology to visualizing and clustering the data of more than 2 variables (Hsu et al., 2009; Hung & Tsai, 2008; Kiang & Fisher, 2008; Wang, 2001). Using the SOM, the data which has high-dimensional structure is visualized in a 2-dimensional space and each data is projected in one of the unit which consists the map (Akay et al., 2010; Kiang & Fisher, 2008). In the map, similar data is located at a near position, and different data is located at a far position; the structure of the data is projected in the graphical map, and it enables a clarified and understandable interpretation of the data. Therefore, the SOM can be used for presenting overviews of high-dimensional data.

In this sbsection, we apply the SOM for visualizing the 4-dimensional (i.e., FRIST, R, F, and M) data and the similarity of the customers. Firstly, customers are divided into the clusters that are decided by the extended RFM analysis based on PLSA model. For the decision of the clusters, the parameters $P(z_k|x_{1i},x_{2i},x_{3i},x_{4i}),(k = 1,\dots,20)$ estimated by the equation (4) are used. For visualizing the data, we assign each customer to the latent class whose probability $P(z_k|x_{1i},x_{2i},x_{3i},x_{4i})$ is the largest. Then we visualize the customer relationships by showing only the mapped data using SOM in the focused latent classes.

Here, we focus on the latent class1 “worst customers” and the latent class 18 “good customers” for the two examples, the visualization of those are shown in Figures 2 and 3.

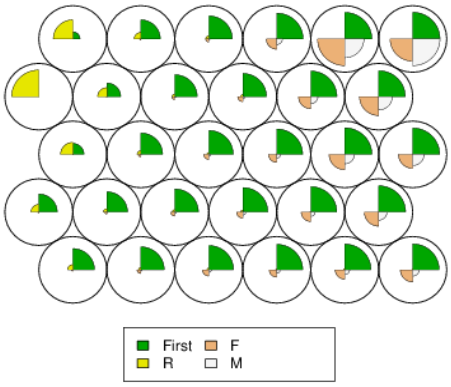

Figure 2: visualization of all data by the SOM.

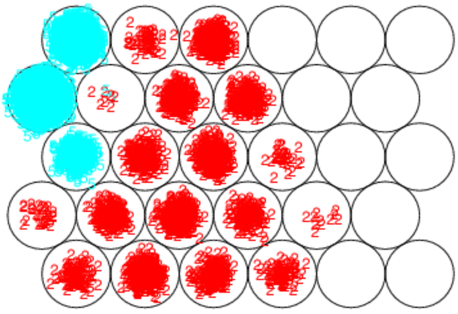

Figure 3: visualization of the latent classes 1 and 18.

In Figure 2, each unit shows the characteristics of individual data. In each unit, we describe the average values of the 4 variables (upper right: FIRST, upper left: R, lower left: F, lower right: M). In Figure 3, customers belonging to the latent class 1 referred to “worst customers (gray)” and customers belonging to the latent class 18 referred to “good customers (black)”, are plotted in each unit. Interpreting Figures 2 and 3, we conclude that the characteristics of the cluster can be visualized by figure. The non-prime customers are plotted in the units on upper left sided unit of Figure 3. Their value of FIRST is smaller than other units, the value of R is larger than others, and both values of F and M are the smallest in the 30 units. Besides, the good customers are plotted in the units on the left middle positions in Figure 3, and their value of FIRST is large, R is small, and F is generally lager than the units on the left. In the same manner, the units on the right of Figure 3 are interpreted as the best customers whose values of F and M are generally lager than others.

From the above discussions, we visualized the result of customer segmentation based on our proposed model using the SOM, so that the characteristics of the latent class could be interpreted transparently.

6. CONCLUSION AND FUTURE WORK

In this paper, we added a new factor “the information of first purchase by the customer” to the three factors of RFM analysis based on PLSA model (Zhang, et al., 2015) which utilizes to analyze history purchase data of customers. The effectiveness of our extended model is clarified by the demonstration using a real data of a major Japanese retail company. The customers are segmented by latent classes, and the characteristics of each segment can be interpreted by the estimated model. According to the demonstration, we clarified that it is possible to segment estranged customers by adding the new factor which is important for a retail store as a factor; therefore, the customers’ interpretation is greatly changed. Moreover, we used the SOM method for the visualization of the result of analysis using the extended RFM analysis, that is, the visualization of the characteristics of latent classes and the similarity of the customers.

A future work is to visualize the customers’ purchase habits (purchase item) of latent classes by the SOM. In addition, it also necessary to involve a segment of new customers into the analysis because our verified result could not employ for new customers but it is important for a retail store.

ACKNOWLEDGMENTS

The authors wish to acknowledge all members of Goto Laboratory, Waseda University, and the Joint Association Study Group of Management Science for their support of our research. This study was partially supported by JSPS KAKENHI, Grant Numbers 26282090 and 26560167.

REFERENCES

Akay, M. F., Abasikeles, I., & Oral, M. (2010) Application of self organizing maps for investigating network latency on a broadcast-based distributed shared memory multiprocessor, *Expert Systems with Applications*, **37**, 2937-2942.

Beane, T. P., & Ennis, D. M. (1987) Market segmentation: a review, *European Journal of Marketing*, **21**, (5), 20-42.

Birant, D. (2011) Data mining using RFM analysis, *Knowledge-Oriented Applications in Data Mining*, 99-108.

Bult, J. R., & Wansbeek, T. (1995) Optimal selection for direct mail, *Marketing Science*, **14**, (4), 378-394.

Hanafizadeh, P., & Mirzazadeh, M. (2011) Visualizing market segmentation using self-organizing maps and fuzzy Delphi method - ADSL market of a telecommunication company, *Expert Systems with Applications*, **38**, 198-205.

Hofmann, T., & Puzicha, J. (1999) Latent class models for collaborative filtering, *Proceedings of the sixteenth International Joint Conference on Artificial Intelligence*, 688-693.

Hsu, S. H., Hsieh, P. A., Chih, T. C., & Hsu, K. C. (2009) A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, *Expert Systems with Applications*, **36**, 7947-7951.

Huang, S., Chang, E. C., & Wu, H. H. (2009) A case study of applying data mining techniques in an outfitter’s customer value analysis, *Expert Systems with Applications*, **36**, 5909-5915.

Hofmann, T. (1999) Probabilistic latent semantic analysis, *Proceedings of the fifteenth International Joint Conference on Uncertainty in Artificial Intelligence*, 289-296.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011) Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study, *Procedia Computer Science*, **3**, 57-63.

Kiang, M. Y., & Fisher, D. M. (2008) Selecting the right MBA schools - An application of self-organizing map networks, *Expert Systems with Applications*, **35**, 946-955.

Tsai, C. Y., & Chiu, C. C. (2004) A purchase-based market segmentation methodology, *Expert Systems with Applications*, **27**, 265-276.

Wang, S. (2001) Cluster analysis using a validated self-organizing method: Cases of problem identification, *International Journal of Intelligent Systems in Accounting, Finance and Management*, **10**, (2), 127-138.

Wei, J. T., Lin, S. Y., Weng, C. C., & Wu, H. H. (2012) A case study of applying LRFM model in market segmentation of a children’s dental clinic, *Expert Systems with Applications*, **39**, 5, 5529-5533.

Zhang, Q., Yamashita, H., Mikawa, K., & Goto, M. (2015) Analysis of purchase history data based on a new latent class model for RFM analysis, *Proceedings of the ASMSA on Management Science and Application*, Dalian.