Data Analysis and Consumption Model Research of E-card system

Liming Xue[1,a], Weixin Luan[2,b]

[1]Department of Management, Dalian Maritime University, Dalian, 116026, China

[2]Department of Management, Dalian Maritime University, Dalian, 116026, China

[a]email: xuelm@dlmu.edu.cn, [b]email:weixinl@dlmu.edu.cn

Keywords: Discretization; Association rule; Decision tree

Abstract. E-card system is widely used in campus. This paper analyses the data of a campus E-card system. First of all, extracted the transaction data from the database and processed the initial data with different methods. For the second, analyze the record amount data with cross table and histogram. For the third, analyze the expense data with data mining methods and compare the difference between different ways of discrete methods.

Introduction

As the development of information technique, many campuses have built their own E-card systems. The E-card system has a large amount of users covering almost every people in campus. According to a median scaled campus with nearly 20000 people, nearly 100000 records were appended every day. Valuable rules and knowledge will be found by processing and analyzing the data. In this paper, we study the E-card system data of a maritime university in china. We choose and design the models and rebuilt the database to make the data adapted to the analysis model. Finally we get some rules which would be useful to the campus management.

Data Extraction and Process

The campus E-card system has more than 20000 users including undergraduates, graduates, doctors, teachers and some temp training students. The system produces nearly one hundred thousand records everyday including restaurant consumption records, shop consumption records, bath consumption records and others. In this paper, we select the shop consumption as the research object.

The transaction database is huge and complicated which owns a lot of tables. Taking the analysis efficiency into consideration, we reduce some tables and fields which would not be useful to analysis result and compose a new table which includes the necessary fields we need from different tables. The structure of new table is as follows:

recordid	cusnumber	custype	sex	shopid	amount	transtime
1	222012xxxx	10	1	21	1.2	063921
2	1997xxxx	14	2	10	4	064109
…	…	…	…	…	…	…
6110	222012xxxx	11	1	90	2.8	222912

In this table, recordid is the primary key, cusnumber is the ID of customer, custype is the ID of customer type, 10 represents undergraduates,11 represents graduates, 14 represents teachers. Sex means the customer sex,1 represents male,2 represents female. Shopid means the ID of shop, amount mean the expense of every transaction, transtime means the transaction time. For the first record as a example, it means a male undergraduate student with a number of 222012xxxx expense 1.2 at No.21 shop in 06:39:21.

There are a lot of algorisms in data mining area such as neural network, association rule, decision tree and others. However, some of them are only adapted to discrete data. When we analyze other data, the algorism such as association rule will not work. In this case, we need to discrete some initial data. The procedure of discrete is important which is closely related to the

result of data mining[1]. In this paper, we discrete the data of transaction time and expense.

Discrete by equal value span

For the convenience of analysis, we add a new field named transtime1 which means the discrete result of transaction time. We discrete the transaction time to time section by half hour. The time before 7:00 is value 1 in transtime1, the transaction time between 7:00 and 7:30 is value 2 in transtime1 and the transaction time between 22:30 and 23:00 is value 33 in transtime1.The matching table is as follows:

transtime	6:30-7:00	7:00-7:30	7:30-8:00	…	21:30-22:00	22:00-22:30	22:30-23:00
transtime1	1	2	3	…	31	32	33

Discrete by equal number of records

Discrete by equal number of records means separate the dataset into different sections by same number of records. In this paper, we have 6110 records in all, if we separate the dataset into 4 sections, every section has about 1500 records. We discrete the dataset as following steps:

(1)Sort all the records by the value of expense (field amount), the new dataset is like {r1,r2,ri,ri+1...r6110}, r1 means the transaction which has the minimum expense, ai means the expense of ri.ai-1<ai<ai+1

(2)Take 1500 as step number,{r1…rm1} belong to the first section, and am1=a1500,am1+1>am1.{rm1+1…rm2} belong to the second section,am2=a3000 and am2+1>am2,go as the same rule.rm1,rm2,rm3 are the points of division.

(3)Add a new field named a-clus1 which means the result of expense which is discrete by equal number of records. For all the records, match the field a-clus1 with field amount.

Discrete by Kohonen neural network cluster

Kohonen neural network is a kind of self-organizing networks, it can recognize the dataset characteristic and cluster automatically[2]. This network adjusts the network weights by self-organizing feature. The network has two layers including input layer and output layer. Neurons connected the two layers and map the input layer on output layer with a two -dimensional discrete graphics. The structure of Kohonen network is as Figure 1:[3]

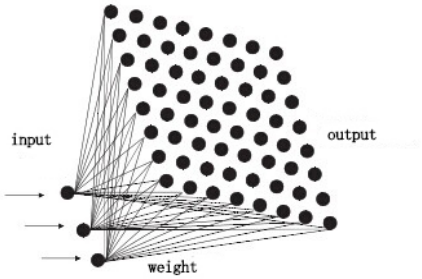


Figure 1

In this case, we cluster the expense data with Kohonen network by a data analysis tool. Set the parameter of learning rate as 0.4 and the cluster number as 4.After calculation, the first cluster covers the expense between 0 and 3.16,the second cluster range from 3.2 to 8.6,the third cluster contains the expense from 8.7 to 15.8 and the expense from 15.9 to 20 formed the fourth cluster.

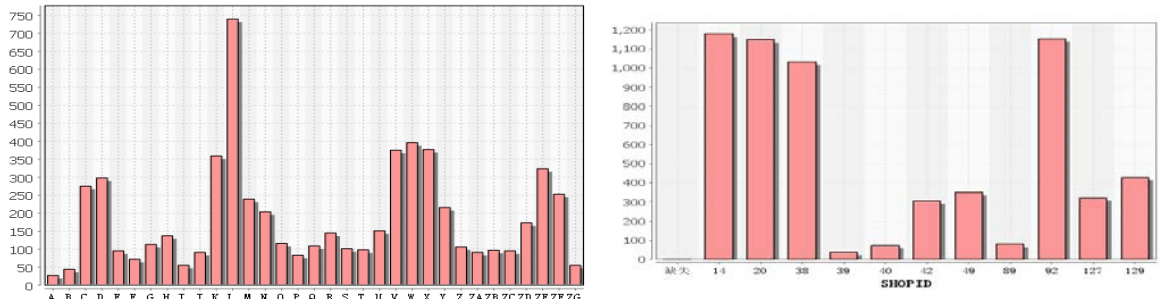
Add a new field named a-clus2 which means the result of expense discrete by Kohonen network. For all the records, match the field a-clus2 with field amount. The matching table is as follows:

amount	a_1	…	a_{1783}	A_{1784}	…	a_{3210}	a_{3211}	…	a_{4528}	a_{4529}	a_{6110}
a-clus1	1	…	1	2	…	2	3	…	3	4	4

Consumption record amount analysis

Analyze based on time-dimension

Take the discrete time as horizontal axis, the amount of records as vertical axis. The Bar graph is as Figure 2 (a):



(a) Based on time (b) Based on location

Fig.2. Consumption record amount

From this figure, we can see that the record amount during 7:30 to 8:30, 11:30 to 13:30, 17:00 to 18:30,21:00 to 22:30 are larger than other time sections, and the time section 12(12:00-12:30) has the largest amount of record.

Analyze based on location

In this case, different shop located in different place. To find the relationship between location and record amount, we choose the shop id as horizontal axis, the record amount as vertical axis. The Bar graph is as Figure 2 (b):

From this figure, we can see that No14, No20, No38, No92 shop have more consumption records than other shops.

Analyze the customer type factor and shop id factor with cross table method, the result is as follows:

	Shop-id	14	20	38	39	…	129	total
Cus10	Count	1087	1051	1014	33	…	301	5371
Cus11	Count	92	96	14	3	…	88	576
…	…	…	…	…	…	…	…	…
Total	Count	79	1148	1032	36	…	429	6110
		19.2931	18.7856	16.8876	0.5891	…	7.0201	

From this table, we can see that in No38 shop, the undergraduates have more than 98% consumption records, In No92 shop, the graduates have about 20% consumption records which is obviously more than other shops. Take location into consideration, we found that the No38 shop near the area of undergraduates and the No 92 near the graduates.

Expense Analysis

Association rule learning is a popular and well researched method for discovering interesting rela tions between variables in large databases. It is intended to identify strong rules discovered in datab ases using different measures of interestingness. The concept of association rules was popularised p articularly due to the 1993 article of Agrawal et al. A famous story about association rule mining is t he "beer and diaper" story. A purported survey of behavior of supermarket shoppers discovered that customers (presumably young men) who buy diapers tend also to buy beer. This anecdote became p opular as an example of how unexpected association rules might be found from everyday data. [4]

In this case, we analyze the consumption time and the expense with association rule. Take the discrete consumption time as input and the expense amount a-clus1 which is discrete by equal number records as object, set the support as 6%, the confidence as 70%. The rules are as follows:

Rule1: consumption time between 12:00 and 12:30, expense between 2 and 3.5, the support is 8.5%, the confidence is 70.6%.

Rule2: consumption time between 17:00 and 17:30, expense between 3.5 and 5.5, the support is 6.1%, the confidence is 100%.

Rule3: consumption time between 17:30 and 18:00, expense between 3.5 and 5.5, the support is 6.5%, the confidence is 100%.

When we take the shop id as the input and the expense amount a-clus2 which is discrete by Kohonen as object, set the support as 5% and the confidence as 70%。The rule is as follows:

Rule1:In No 38 shop, the expense less than 3.16,the support is 11.8% , the confidence is 70.6%.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. There are 3 types of nodes in a decision tree: (1) Decision nodes - commonly represented by squares; (2) Chance nodes - represented by circles; (3) End nodes - represented by triangles. Among decision support tools, decision trees have several advantages, i.e., simple to understand and interpret, allow the addition of new possible scenarios, can be combined with other decision techniques. [5]

In this case, take the discrete expense a-clus1,a-clus2 as object separately, take the consumer type and consumer sex as input, set the confidence as 50%, the results are as the following table：

No	object	rule
1	a-clus1	Undergraduate a-clus1=4（5.5-20）
2	a-clus1	Doctor a-clus1=4（5.5-20）
3	a-clus1	Teacher a-clus1=4（5.5-20）
4	a-clus1	Training a-clus1=4（5.5-20）
5	a-clus2	Undergraduate a-clus2=1（0.1-3.16）
6	a-clus2	Graduate a-clus2=3（8.7-15.8）
7	a-clus2	Doctor a-clus2=3（8.7-15.8）
8	a-clus2	Teacher a-clus2=3（8.7-15.8）
9	a-clus2	Training a-clus2=4（15.9-20）

Comparing with the consumer type, the consumer sex has nearly no influence. When we take a-clus1 as object, we found that the consumer except undergraduates has the expense between 5.5 and 20.When we take a-clus2 as object, we found that the undergraduates have the expense less than 3.16, the graduates doctors and teachers have the expense between 8.7 and 15.8, the training students have the expense between 15.9 and 20.

Conclusion

This paper studies the data of E-card system, analyze the time, expense, consumer data with data mining methods such as decision tree, Kohonen network, association rule. Get some rules and conclusions which will be useful to support the management and decision of campus.

References

[1] Yu Sang: Research on Discretization Methods for Continuous Data; Dalian University of Technology 2012

[2] Chunping Liu: Comparison of clustering methods based on Kohonen neural network

in remote sensing classification[J];Journal of Computer Applications 2006.26(7)1744-1746.

[3] MA Shu-qin :Research on network intrusion clustering based on Kohonen neural network algorithm; CHINA MEASUREMENT & TEST 2013.39（4）113-117

[4]Piatetsky-Shapiro, Gregory (1991),Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.Agrawal, R.; Imieliński, T.; 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 - SIGMOD '93". p. 207.

[5] Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies 27 (3): 221.