Studies and Applies of the Data Mining in the Mathematical Modelling

He Liu, XingGang Zhang
Department of Computer and Science,
Cheng Dong of Northeast Agricultural University
Harbin,150025,China
e-mail: 491084313@qq.com

***Abstract*—Along with the coming of information age, many websites use its formidable resources and popularity, provides day by day the service of specialized and convenient to its member group. The members online rent DVD, became the website management administrative personnel the issue of concern. How to having the data carries on the analysis processing, is solves this problem the key. The article through building the mathematical model, has solved this problem using the data mining technique very well.**

Keywords-Data mining; Rational distribution; Degree of satisfaction

I. Introduction

Along with the maturity and popularization of data of application data bank technology, the data quantity that humanity accumulates rapidly is growing by the index speed. Unfolds in front of the people have not limited to this vast databank of department, this unit and this profession, but is the vast boundless information sea. Faced with vast boundless data, people summoned from the data vastly the technology that came one to discard the dross and select the essential and eliminate the false and keep the true, discovered the knowledge and core technologies from the database. The data mining then arose at the historic moment. The KDD from the data discovered that useful knowledge entire process, the data mining is one specific step in KDD process, is the KDD core, it uses the decimation mode of special algorithm from the data. This pattern is new, possibly useful and is understandable finally. The so-called data mining, extracts concealed from the database, beforehand unknown, has potential application value the process of information. What the data mining and traditional analysis tool are different, data mining use bases on the method of discovery, essential links between other use of models matches and algorithm determination data. The quality of data mining algorithm will affect directly to discover the knowledge the application degree. It is the process of relapse, usually contains many steps of interconnections: Pretreats and comes up with the hypothesis, selection algorithm, extraction rule, appraisal and explanation result, the pattern constitution knowledge and application[1-2].

The following some year national university student mathematics modeling contest looks at the practical application of data mining for the example

II. Related Work

A. Problem analysis

Online rents DVD, how online rents the DVD 3 small issues is to the different type DVD quantity should assign, can achieve to need the effect. And the issue 1 to meet one part of members' requirements, prepares 5 kinds of different types the DVD number of sheets; The issue 2 in the situations of known 20 kinds of DVD existing quantities, how assigns them according to being partial degree of member, enabling the degree of satisfaction of member to achieve in a big way; Issue 3 and issue 2 similar, but the manager of request in one month purchases many 20 kinds of different types DVD, can cause DVD that one part of members can see to expect. Regarding the issue 1, according to 1000 members who in Table 1 investigates, is willing to watch the type 1, type 2, type 3, type 4, type 5 population to come to speculate ideally in 100,000 members is willing to watch these 5 kinds of DVD population, may know by the supposition, in these 5 kinds of members not overlapped personnel appear. The historical data showed that 60% members rent every month twice, 40% only rents one time. Rents DVD to be two to 60% members, we can transform are DVD are rented by two different members successively, is the population that it satisfies for them. These two parts of sums are to hope to see that in this DVD member at least one month of 50% in the member number that in can see. For the rationality of more precise explanation model, from probability, builds a model, these two models confirm mutually. In the time is three months, establishes 95% to see the model and this model that this DVD member counts are similar. Regarding issue 2, because in the topic does not have the explanation in other in January or cycles, unifies being partial degree of member, is partial to the degree with its online order digit (1-9) being in reverse proportion example, the administrative personnel existing DVD to provide to the member one time, sought at this time overall greatest degree of satisfaction, the degree of satisfaction with being partial to the degree had certain relations. Regarding issue 3, what we consider is 95% assignment problems. Has one part of members to obtain 6 kinds in January, some obtain 3 kinds, but 5% members have not obtained DVD that any kind of he hopes[3-4].

Published by Atlantis Press, Paris, France.

B. *hypothesis of model*

1)express company transmits one time to require the time is 4- 6 days, because the member in company resides in land, then we take for 5 days are its average express time.

2) suppositions receive the DVD member to look completely them in three days, and looked, ensure through express mails back promptly completely, like this conforms to one part of members to rent DVD in January to be two, one part of members rent DVD one time.

3) DVD that the websites after receiving the member mail back, distributes to other members promptly

4) the data probability of 60% and 40% to each kind of DVD and each member are equal

5)Assuming that each DVD unit price is the same

6) in the questionnaire survey, is willing to watch the different type DVD personnel is not overlapped unusualness carries on

7) In from one to three months the member quantity is invariable

8)Does not have the turnover order that other reasons (for example network interrupt) causes carries on unusually[5]

C. *nomenclature*

A1: Some website existing member counts

A2: The member of present need processing counts

A3: Member of questionnaire survey counts

Bj: Is willing to watch j (j= 1, 2, 3, 4 and 5) the kind of DVD member to count

b: Every month rents the proportion of DVD time member. And a+ b= 1

bi: 0, 1 variable, when i= 0 has not obtained DVD, when i= one obtains DVD

K: Every month rents DVD is one time

H: Can obtain the hope to see this DVD the proportion of member

Xj: In one month ensure at least 50% members see the hope DVD, j (j= 1, 2, 3, 4 and 5) the DVD quantity

Yj: In three months to guarantee at least 95% members may see DVD that hopes, needs j (j= 1, 2, 3, 4 and 5) the kind of DVD quantity

xij: The ith member obtains the serial number is j DVD, belongs to 0 and 1 variable. (I= 0001 and 0002,0100, j= 001 and 002, 020)

dij: The ith member to the serial number is j being partial of DV

Pj: The serial number is the j DVD existing quantity. (J= 001 and 002, 020) Nj: The serial number is the j DVD purchasing volume. (J= 001 and 002, 020)

5 model forming and solution[6]

III. EXPERIMENT ANALYSIS AND RESULT

The existing DVD quantity and member order, assigns, transmits and returns to five relations be possible the image to express is:

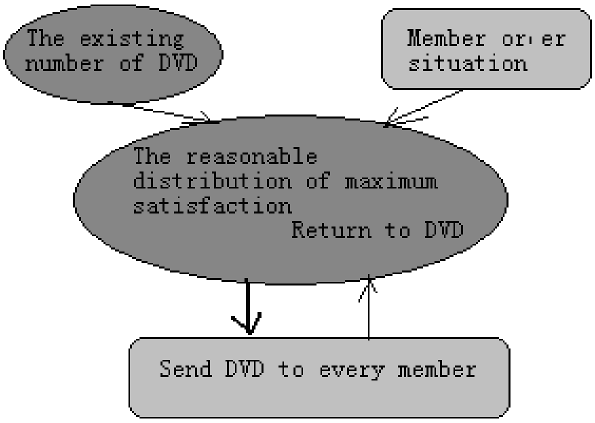
The existing number of DVD
Member order situation
The reasonable distribution of maximum satisfaction
Return to DVD
Send DVD to every member

Figure 1 Five between relations

Question 1. The problem is in order to achieve the requirements of part of the members shall be prepared at least how many pieces of different kinds of DVD. The analysis can reach the conclusion that model:

$$X_j \times a \times 2\lambda + X_j \times b \times \lambda = A_1 \times \theta \times \tfrac{B_j}{A_3} \qquad \ldots\ldots (1)$$

According to Table 1:

$A_1 = 100000$; $A_3 = 1000$; $B_1 = 200$; $B_2 = 100$; $B_3 = 50$; $B_4 = 25$; $B_5 = 10$; $\lambda = 1$; $\theta = 50\%$; $a = 60\%; b = 40\%$ Substitution

$$X_1 \times 0.6 \times 2 + X_1 \times 0.4 = 100000 \times 0.5 \times 200 \div 1000$$

$$X_2 \times 0.6 \times 2 + X_2 \times 0.4 = 100000 \times 0.5 \times 100 \div 1000$$

$$X_3 \times 0.6 \times 2 + X_3 \times 0.4 = 100000 \times 0.5 \times 50 \div 1000$$

$$X_4 \times 0.6 \times 2 + X_4 \times 0.4 = 100000 \times 0.5 \times 25 \div 1000$$

$$X_5 \times 0.6 \times 2 + X_5 \times 0.4 = 100000 \times 0.5 \times 10 \div 1000$$

With MATLAB software solution: Five kinds of DVD numbers of sheets respectively are

$X_1 = 6250$, $X_2 = 3125$, $X_3 = 1563$, $X_4 = 781$, $X_5 = 313$

Likewise, the models within three month are:

$$Y_j \times a \times 6\lambda + Y_j \times b \times 3\lambda = A_1 \times \theta \times \tfrac{B_j}{A_3} \qquad \ldots (2)$$

$\theta = 95\%$ substitute, other data with on

$$Y_1 \times 0.6 \times 6 + Y_1 \times 0.4 \times 3 = 100000 \times 0.95 \times 200 \div 1000$$

$$Y_2 \times 0.6 \times 6 + Y_2 \times 0.4 \times 3 = 100000 \times 0.95 \times 100 \div 1000$$

$$Y_3 \times 0.6 \times 6 + Y_3 \times 0.4 \times 3 = 100000 \times 0.95 \times 50 \div 1000$$

$$Y_4 \times 0.6 \times 6 + Y_4 \times 0.4 \times 3 = 100000 \times 0.95 \times 25 \div 1000$$

$$Y_5 \times 0.6 \times 6 + Y_5 \times 0.4 \times 3 = 100000 \times 0.95 \times 10 \div 1000$$

With MATLAB software solution: Five kinds of DVD numbers of sheets respectively are

$Y_1 = 3958, Y_2 = 1979, Y_3 = 990, Y_4 = 495, Y_5 = 198$

With theory of probability to primary model verification:

Events B_1 for each monthly rent two members $P(B_1) = 0.6$

Events B_2 for each monthly rent a member $P(B_2) = 0.4$

The event A is can get hope required DVD member $P(A_1) \geq 0.5$ $P(A_2) \geq 0.95$

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2) \qquad \ldots\ldots (3)$$

Substituting data

$$\frac{x_1 \times 0.6}{200 \times 100 \times 0.6} \times 0.6 + \frac{x_1}{200 \times 100 \times 0.4} \times 0.4 = 0.5$$

$$\frac{x_2 \times 0.6}{100 \times 100 \times 0.6} \times 0.6 + \frac{x_2}{100 \times 100 \times 0.4} \times 0.4 = 0.5$$

$$\frac{x_3 \times 0.6}{50 \times 100 \times 0.6} \times 0.6 + \frac{x_3}{50 \times 100 \times 0.4} \times 0.4 = 0.5$$

$$\frac{x_4 \times 0.6}{25 \times 100 \times 0.6} \times 0.6 + \frac{x_4}{25 \times 100 \times 0.4} \times 0.4 = 0.5$$

$$\frac{x_5 \times 0.6}{10 \times 100 \times 0.6} \times 0.6 + \frac{x_5}{10 \times 100 \times 0.4} \times 0.4 = 0.5$$

That same

$X_1 = 6250$, $X_2 = 3125$, $X_3 = 1563$, $X_4 = 781$, $X_5 = 313$

Three months similar:

This model is in the ideal state, which is, namely, in random investigation without overlapping personnel appear, according to the survey launched. The number in more cases, this proportion is not accurate, so will influence in preparation for DVD. In order to improve the accuracy, we should survey the concrete analysis.

Question 2, according to the problem analysis and model hypothesis, the establishment of the model :

The objective function:

$$MaxY = \sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} d_{ij} \qquad \ldots\ldots (4)$$

Constraint conditions:

$$\sum_{j=1}^{20} x_{1j} = 3 \quad \ldots\ldots \cdots \ldots\ldots \sum_{j=1}^{20} x_{100j} = 3$$

$$\sum_{i=1}^{100} x_{i1} = P_1 \quad \ldots\ldots \cdots \ldots\ldots \sum_{i=1}^{100} x_{i20} = P_{20}$$

Where

$P_1 = 8$, $P_2 = 1$, $P_3 = 22$, $P_4 = 10$, $P_5 = 8$, $P_6 = 40$, $P_7 = 1$, $P_8 = 1$, $P_9 = 8$, $P_{10} = 15$, $P_{11} = 19$, $P_{12} = 20$, $P_{13} = 10$, $P_{14} = 2$, $P_{15} = 5$, $P_{16} = 8$, $P_{17} = 30$, $P_{18} = 10$, $P_{19} = 8$, $P_{20} = 38$

Using LINGO software to calculate the degree of satisfaction for 152.99, can get first 30 members have three kinds of DVD case.

Question 3. 95% of the members within one month get he want to see DVD us to the 95% can be divided into three kinds of cases (as shown in figure 1 shows) discussion, and the satisfaction of the third case of each DVD number approximatively as the first and the second kind of average, we ask also is approximatively as for third scheme.

1) 95% for all rent two times, and then were to DVD number of sheets for: 100 * 3 * 95% = 285 copies

2) each one rent a, the common need to DVD number of sheets for: 100 * 95% * 3/2 = 142.5 zhang, rounding off for 143 zhang

3) 95% of the people, including the lease one and lease 2 times

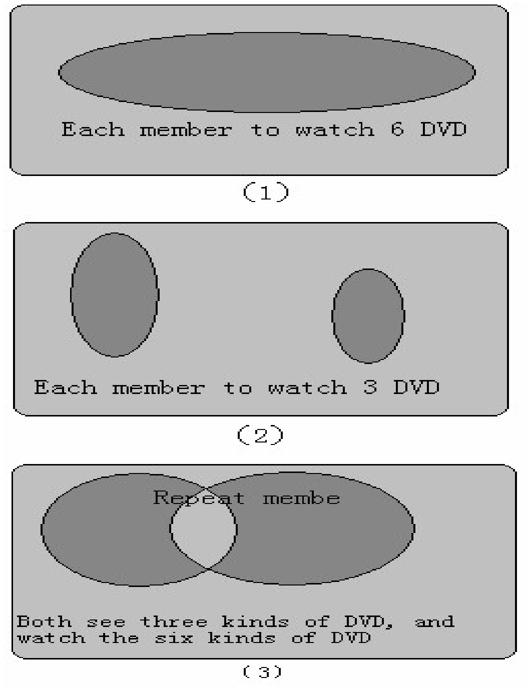
Each member to watch 6 DVD
(1)
Each member to watch 3 DVD
(2)
Repeat membe
Both see three kinds of DVD, and watch the six kinds of DVD
(3)

Figure 2 Circle said 1 set, one or two additive for 95%, rectangular said to 1

If using integer said operation quantity more big. We can think of a number decimal part in less than a certain value, the decimal part take out, then in the nearby obtained integral value, and then return back to model generation can be obtained by the approximate optimal value.

In the first kind of circumstance:

The objective function:

$$MaxY = b_i \sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} d_{ij} \qquad \ldots\ldots (5)$$

Constraint conditions

$$\sum_{j=1}^{20} x_{1j} = 3 \quad \ldots\ldots \cdots \ldots\ldots \sum_{j=1}^{20} x_{100j} = 3$$

$$\sum_{i=1}^{100} x_{i1} = 2P_1 \quad \ldots\ldots \cdots \ldots\ldots \sum_{i=1}^{100} x_{i20} = 2P_{20}$$

$$\sum_{j=1}^{20} P_j = 285 \qquad \sum_{i=1}^{100} b_i = 95$$

By using LINGO software to calculate the maximum satisfaction is 232.750

Where $b_{75}, b_{93}, b_{94}, b_{97}, b_{99}=0$

$P_1=16.5$, $P_2=11$, $P_3=15$, $P_4=13$, $P_5=12.5$, $P_6=15$, $P_7=13.5$, $P_8=18$, $P_9=13.5$, $P_{10}=15.5$, $P_{11}=15.5$, $P_{12}=14$, $P_{13}=15$, $P_{14}=13$, $P_{15}=16$, $P_{16}=13.5$, $P_{17}=14$, $P_{18}=15$, $P_{19}=14.5$, $P_{20}=11$

Then return and the original model: less than 15 number if there is a decimal, then into for integer, more than 15 of the number has the decimal when is to go back for integer, then we get the optimal value of 232.7

In order to reduce error, we once again put less than 15 number if there is a decimal, is to go back for integer, more than 15 of the number has the decimal by into for integer, we get the optimal value of 232.8

In the second case:

The objective function:

$$TMaxY = b_i \sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} d_{ij} \qquad (6)$$

Constraint conditions

$$\sum_{j=1}^{20} x_{1j} = 3b_1 \ \ldots\ldots \ \ldots \ \ldots\ldots \sum_{j=1}^{20} x_{100j} = 3b_{20}$$

$$\sum_{i=1}^{100} x_{i1} = 2P_1 \ \ldots \sum_{i=1}^{100} x_{i20} = 20P_{20}$$

$$\sum_{j=1}^{20} P_j = 133 \quad \sum_{i=1}^{100} b_i = 48$$

By using LINGO software to calculate the maximum satisfaction is 174.1635

Where

$P_1=9$, $P_2=5$, $P_3=7.5$, $P_4=7$, $P_5=5.5$, $P_6=9.5$, $P_7=7$, $P_8=8.5$, $P_9=9$, $P_{10}=8$, $P_{11}=7$, $P_{12}=8$, $P_{13}=8$, $P_{14}=5.5$, $P_{15}=7.5$, $P_{16}=5$, $P_{17}=7$, $P_{18}=7.5$, $P_{19}=8.5$, $P_{20}=2.5$

Then return and the original model: less than 6 number if there is a decimal, then into for integer, greater than 6 number have decimal when is to go back for integer %, we get the optimal value of 174.1768.

In order to reduce error, we once again put less than 6 number if there is a decimal, is to go back for integer, greater than 6 number have decimal by into for integer, we get the optimal value of 175.2135

finally, we take the two cases of average, get

In the third case:

$Y = (232.760 + 232.860 + 174.1768 + 175.2135) / 4 = 203.7526$

$P_1=14$, $P_2=10$, $P_3=10$, $P_4=10$, $P_5=9$, $P_6=12$, $P_7=12$, P_8=13 , $P_9=12$, $P_{10}=15$, $P_{11}=18$, $P_{12}=16$, $P_{13}=21$, $P_{14}=16$, $P_{15}=15$, $P_{16}=18$, $P_{17}=26$, $P_{18}=15$, $P_{19}=15$, $P_{20}=8$

We solve the problem with three approximation algorithm, this algorithm inevitably exist the bigger error, so we plan to three improvement:

hypothesis $MaxY = \sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} d_{ij} \qquad P = \sum_{j=1}^{20} P_j$

The objective function: $MaxM = \dfrac{Y}{P}$

$$\sum_{i=1}^{100} x_{i1} \le 2P_1 \ \ldots\ldots \ \ldots\ldots \ \ldots\ldots \sum_{i=1}^{100} x_{i20} \le 2P_{20}$$

$$\sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} = \left| 3\left(\sum_{i=1}^{100} \sum_{j=1}^{20} x_{ij} - 4\right) \times b_i \right| \qquad \sum_{i=1}^{100} b_i = 95$$

IV. CONCLUSION:

Promotes this kind of model in other aspects serviceability.For example, information high-speed developed today, bank loan issue and investment issue. This kind of issue data quantity is getting bigger and bigger, we must to the existing data statistics analysis. The consumers can act according to their ability to determine to their benefit biggest that one kind of loan, the factory to enhance own popularity and prestige, also from own benefit, how to estimate and fixed purchase commodity through the network, how the postage determines and so on a series of issues. The application in travel industry, which one day arranges many people to travel to which place, can achieve according to the model both enhanced the status in audience mind, and improved to develop itself.

REFERENCES

[1] XU Cong.data mining method application in traditional forecast model.Hebei Journal of Industrial Science and Technology 2009,pp.120-108

[2] LI Dong-ping.MING Shi-xiang.Application of data mining in mine production management domain.Mineral engineering,2006,pp.56-62

[3] Yang jiaxu.Data warehouse-based mathematical model solution the research and designs.Tongji University Software Institute,2007,pp.78—82

[4] Liu Xue.KDD and data mining in mathema tical modelling.China technology information,2006,pp.39-43

[5] Huang Ke-kun .Purchase and apportion model that DVD online rents . Institute journal, 2007, 34-40

[6] DU Ya-nan,SHAO Guang-li.DVD online rents the issue the optimal solution.Central University of Nationalities journal (natural sciences version),2007,PP.16-22