American Journal of Industrial and Business Management, 2015, 5, 813-818
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ajibm
http://dx.doi.org/10.4236/ajibm.2015.512078

A Study on Customer Segmentation for E-Commerce Using the Generalized Association Rules and Decision Tree

Haiying Ma

Department of Management Science and Engineering, East China University of Science & Technology, Shanghai, China
Email: c_mhy@163.com

Received 21 November 2015; accepted 21 December 2015; published 24 December 2015

Abstract

With the rapid development of e-commerce, e-commerce is becoming more and more competitive. How to improve customer loyalty, attract more new customers, and expand the market effectively, it is very important for the e-commerce enterprise. In this paper, a comprehensive model is proposed, which is based on generalized association rules and decision tree technology. The model is used for customer segmentation of e-commerce website. It can help e-commerce companies understand customers, support decision-making, so as to provide customers with more targeted services.

Keywords

Customer Segmentation, Association Rules, Decision Tree

1. Introduction

With the rapid development of science and technology, the Internet is playing a more and more important role in people's life, study and work, thus leading to the increasing growth of e-commerce and competition. To win the competition, e-commerce businesses have to project effectively into the potential purchase of the customer, offer customized service and make relevant marketing strategies [1].

With the continuous development of e-commerce, the traditional technique of customer segmentation has been unable to cope with the massive and complex customer data. Based on the data mining technique, the new analyzing technique provides new solutions to the massive data of complex customer segmentation. Through collecting and classifying customer information, the new technique intends to find out customer groups with

How to cite this paper: Ma, H.Y. (2015) A Study on Customer Segmentation for E-Commerce Using the Generalized Association Rules and Decision Tree. *American Journal of Industrial and Business Management*, **5**, 813-818.
http://dx.doi.org/10.4236/ajibm.2015.512078

different attribute features: the demand characteristics of the overall customer internal, the buying behavior, the browsing characteristics and etc. Then it subdivides customers, helps e-commerce businesses understand their customers, provides clustering customer groups with more suitable, comprehensive and customized service, selects the most exploitable target customer groups and finds out the most potential customers.

2. Research Method

When evaluating a model, we usually put three things into consideration: (1) forecast accuracy; (2) stability; (3) interpretability [2]. Out of question, forecasting accuracy is the primary consideration of the modelers. But stability may be more important in China, where the consumer market of B2C e-commerce is developing rapidly. As the e-commerce market is developing rapidly, there may be immediate differences arising between all of the newly-joined customers and the overall modeling where a more stable model will be more popular with its users. Thus, the key problem lies in whether we can build a better model by integrating different techniques and applying all means of technical features.

Clustering, neural network and some of the other techniques are commonly used for customer segmentation. However, they all have their own defects.

When there are doubts about the natural grouping, we can apply the clustering technique to represent the numerous common customer groups. During the process of customer segmentation, the clustering technique is more comprehensible than most of the other ones. It is a kind of unsupervised technique and does not require relevant prior knowledge. However, the outputs obtained with the clustering algorithm cannot explain themselves, only to be comprehended by other techniques [3]. This insufficiency can be overcome by applying the integration of the clustering and decision tree techniques. Generalized association rule model overcomes the shortcomings of clustering, which can be carried out on the multi-level concept level, and the content of the process is more easy to understand. Decision tree has the advantages of high accuracy, simple and efficient, which can not only deal with the “income”, “age” and other numerical data, but also deal with the “gender”, “occupation” and other non numerical data, so it is very suitable for B2C e-commerce website customer segmentation.

The interpretation of model output is very important. Because the output of the model is used to guide the decision of the enterprise [4]. In the process of classifying and forecasting models, the neural network technique is a good choice if obtaining the output is more important than understanding the working principle. The advantage of neural network lies in its better adjustability to the noise data and its better forecasting ability to the unknown data. When there are hundreds of characteristic quantities to input, the effect of the neural network will not be good enough. This insufficiency can be overcome with the integration of neural network and decision tree.

Thus, it can be seen that it is very effective to apply the integrated technique in the customer segmentation model.

3. Modeling

The model of generalized association rule overcomes the shortcomings of clustering and etc. It can be performed on the multilevel-concept basis and can handle broader contents, making rule sets more comprehensible, while there are no restrictions on a single output field [5].

Decision tree has advantages of high accuracy, simplicity, efficiency and etc. It can not only deal with numerical data concerning “income”, “age”, but also deal with other non-numerical data about “gender”, “occupation”. Therefore, decision tree is very suitable for the application of customer segmentation on the B2C e-commerce website.

Such technical integration not only ensures the forecast accuracy of models but also guarantees the stability and interpretability of models. Therefore, we can foresee that it is feasible to integrate the two techniques and apply it to build models during the process of e-commerce customer segmentation.

Specific ideas are as follows: first, build a model of customer segmentation by applying the generalized association rules, analyze the connections between different purchase items to determine customer groups. Then induce rules of demographic features with different customer groups by using the outputs obtained from the generalized association rules and the decision tree model.

3.1. Selection of Model Variables

Customer segmentation is based on the selection of segmentation variables. Generally, there are two types of segmentation variables-descriptive variables and behavioral variables [6]. Based on the variables of shopping basket model in the traditional supermarket (The index of the traditional model is the classic index, which is correct and reliable) and the actual situation of e-commerce websites. In this paper, we chose two kinds of indexes, such as descriptive variables and behavioral variables. Variables used in the model are as following:

(1) Descriptive variables: descriptive variables of the customer are mainly used to comprehend the basic attributes of customer information. Here are some of the main variables to select: register ID: the registered accounts of website members, sex, age, career, income.

Such indexes play a key role in determining the members of a particular market segment. These variables mainly come from the registered information of members' and the basic information collected through the management system of e-commerce websites.

Such variables are mostly static data, describing the basic attributes of the member. Its advantage consists in that most of the variables are easy to collect. But sometimes basic variables of member description lack differentiation. Some of the variables are often related to member privacy, such as the residence of the member, contact information, income information and etc.

The accuracy of data collection is the leading evaluative feature of the member description variables.

(2) Behavioral variables: behavioral variables mainly refer to a series of variable indexes relating to the connection of member businesses and e-commerce websites.

(3) The main variables are as follows: value (total spending of the member on this site), p method (payment of the member), buy record (record of the purchased service or products), view record (record of the browsed commodities).

Such indexes are used to define where the e-commerce website should strive in a segment market. And they are the key factors in determining the target market. On the e-commerce website with complete systems of member information collection and management, the records of member transaction are easy to attain and are usually perfect from the perspective of transaction records. But what needs to be aware of is that behavioral variables the member are not exactly the same as the records of member transaction and consumption. To attain the behavioral feature of the member, the record of member transactions and other behavioral data have to be processed and analyzed. Then information can be counted and extracted.

Completeness is the main factor in measuring the behavioral information of the member.

3.2. Building of the Integrated Model

The integrated model can be seen as a two-stage model consisting of generalized association rules and decision tree algorithm. Within the integrated model, the outputs of generalized association rules become the rules of potential segmentation for customers on the e-commerce website. Therefore, the algorithm out of generalized association rules is defined as the first stage in the integrated model and the outputs of the first stage serves as the collection of data sample of the decision tree model. Specific steps are as follows:

Step 1: The first stage of the model is to select the variety variable of the purchased commodities from all variables, forming a data item set. Each data within the set corresponds to one type of commodity, constituting a set of objects. Calculate the degree of support for all possible rules: The support of rule X => Y in the data set is the ratio of numbers between data sets with X, Y and all arrangements.

Step 2: On the base of mining, we specify a minimum support, find out all the specified item sets with the minimum support from the database of commodity variety, known as the frequent item sets. Generate the required generalized association rules by applying the frequent item sets.

For example, for frequent item set ABCD and AB, if the ratio conf = support (ABCD)/support (AB) > minconf, the generalized association rule AB => CD will be generated.

Step 3: Trim the non-interest rules off the generalized association rule set. Some purchasing connections will finally come out of the commodity categories. Then apply the connections with customer segmentation rules.

Step 4: On the second stage of the model, decision tree C5.0 can be used to add up and induce the features obtained out of the association rules.

To analyze data by applying decision tree C5.0, the most important step is the tree building process. The generation of decision tree is the process of generating the data item sets by applying generalized association rules on the first stage. During the tree building process, determining the base of testing attributes is very important. Information gain method is used here to help identify the desired testing attributes while each node is generating.

Thus, the attributes with the greatest information gain (the maximum degree of entropy reduction) can be selected to function as the testing attribute of the current node. Using the attribute to divide the current (node contained) sample set will minimize "degree of mixing with different categories" from all obtained sample sets.

Step 5: When a decision tree has just been built, many of its branches are made up of the abnormal data from the training sample set. Pruning the decision tree is needed to prevent the over-fitting of the newly-built tree and the training sample to improve the speed and accuracy of the future classification. To promote the speed and accuracy of classification, pruning is usually carried out to delete the most unreliable branches by means of statistics. We usually adopt two types of pruning, pre-pruning and post-pruning.

Step 6: Extract rules from the pruned decision tree. The categorized knowledge represented by decision tree may be extracted and expressed through the rule of classification format if-then. The if-then format is more comprehensible and its advantage will be more outstanding when the decision tree is relatively big.

A classification rule is made up of a path between the root node of the decision tree and any of its leaf nodes. The "attribute-value" generated along the path of the decision tree makes up the conditional part (if part) of the classification rule, while the categories marked by the leaf node become the conclusion of the rule (then part).

The final conclusions drawn from the above steps are the rules for customer segmentation of the B2C e-commerce website, through which various customer groups can be defined, thus achieving the purpose of customer segmentation.

4. Analysis and Discussions

This article uses Clementine as a tool to create models. Clementine is the platform of data mining tools and it enables the application of the integrated models, which are consistent with the requirements of an integrated customer segmentation model on the e-commerce website.

4.1. Establishing Data

Questionnaires with the intended contents are handed out to 200 members of a certain e-commerce website and 171 valid questionnaires are received.

According to data collected from the questionnaires, we select some to perform model training and testing. First, commodity variables are converted into the form of T-F, T for the purchased, F for the not purchased. See **Figure 1**.

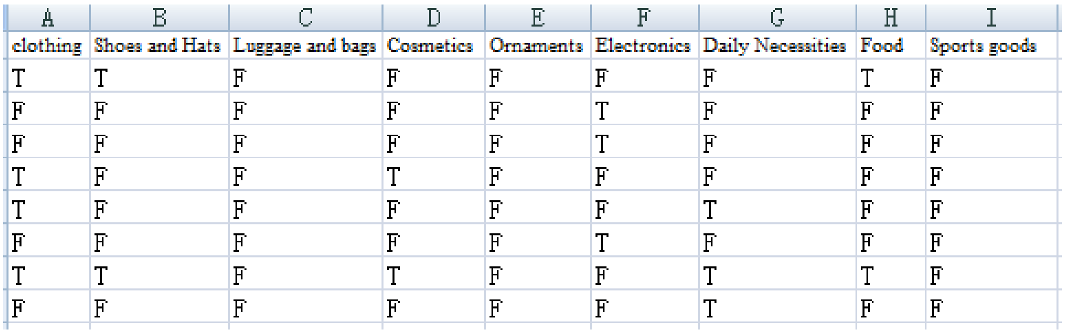
4.2. Clementine Modeling

First, import data from Excel.

Construct a model for the generalized association rules (GRI). Select the item of the purchased commodity as the fields of modeling. Set the direction of fields. Implement the whole GRI nodes and an unrefined model will be generated.

The model is embedded with association rules. Click Browse to observe and discover the rules.

To highlight certain patterns of the data, the network display technique is applied to express the GRI model. Edit the network nodes before performing them. See **Figure 2**. The correlation of commodities can clearly be seen from the network graph, where each strong connection represents a customer group.



A	B	C	D	E	F	G	H	I
clothing	Shoes and Hats	Luggage and bags	Cosmetics	Ornaments	Electronics	Daily Necessities	Food	Sports goods
T	T	F	F	F	F	F	T	F
F	F	F	F	F	T	F	F	F
F	F	F	F	F	T	F	F	F
T	F	F	T	F	F	F	F	F
T	F	F	F	F	F	T	F	F
F	F	F	F	F	T	F	F	F
T	T	F	T	F	F	T	T	F
F	F	F	F	F	F	T	F	F

Figure 1. Model data.

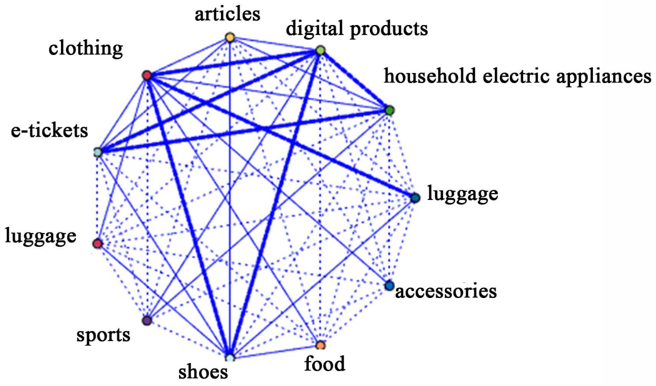


Figure 2. The network graph of association rules.

To induce the demographic characteristics of each customer group, we select strong association rules from the network graph, generating and exporting nodes.

To obtain the demographic characteristics of different customer groups, we set the appropriate field direction for each node to be exported and add a node of decision tree C5.0 respectively. Set the output type as "rule set" and implement the nodes of the decision tree C5.0 respectively.

4.3. Outputs of the Model

Figure 3 is the rule set generated by analyzing the association rules of decision tree C5.0 (clothing and cosmetics as examples). The purpose of customer segmentation is achieved by integrating these rule sets.

4.4. Analysis of the Model Outputs

Results shows the outputs of analysis based on the above three outputs generated out of the integrated model. The accuracy ratios of the three rule sets are 87.13%, 95.91%, 85.96% with the integrated model and 85.38%, 93.57%, 82.46% with the decision tree C5.0 model, shown in **Figure 3**. Thus, the outputs confirm the feasibility of the proposed model. As can be seen from the results, although the accuracy of the output has increased, but the magnitude of the increase is not great. Because the number of samples collected is limited, resulting in the intensity of training is not high. If the number of samples is large enough, the accuracy will be very obvious.

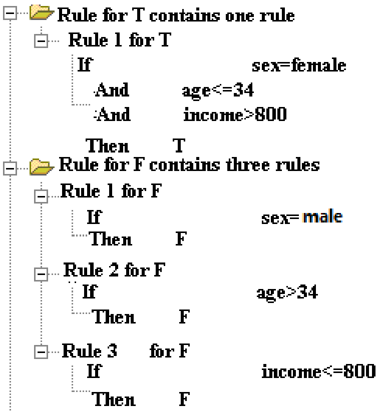
5. Conclusions

As the competition between e-commerce businesses has reached a stage full of conflicts, the problem of customer segmentation is getting more and more attention. Since more operators of e-commerce business are seeking better strategic countermeasures with the help of customer segmentation model, the techniques of customer segmentation have widely been used in different areas of the e-commerce application field. In the past few decades, the application of customer segmentation has been one of the most concerned problems. Modeling techniques, such as the traditional technique of statistic analysis and artificial intelligence, have been greatly developed to achieve the task of successful customer segmentation.

Through analysis and comparison of the traditional model, this paper proposes to build a new integrated customer segmentation model on the e-commerce website: First, build models out of generalized association rules. Then, set the outputs out of the generalized association rules as the algorithm condition of the decision tree. Next, establish the decision tree model. Finally, subdivide customers by applying the decision tree model.

The advantages of establishing the integrated model lie in: The model can develop the technical advantages of the generalized association rules as well as the decision tree. Meanwhile, the integration of the two models can overcome their own shortcomings, making the model optimal. The shortage of the model is that it needs the high accuracy of the training sample; otherwise it is difficult to achieve the expected accuracy. The model needs to be improved in fault tolerance.

Therefore, putting such a model into practice is actually more effective than simply applying one technique to build models.



```
Rule for T contains one rule
  Rule 1 for T
    If                sex=female
     And       age<=34
     And       income>800
    Then      T
Rule for F contains three rules
  Rule 1 for F
    If                sex= male
    Then      F
  Rule 2 for F
    If                age>34
    Then      F
  Rule 3      for F
    If                income<=800
    Then      F
```

Figure 3. Rule sets of clothing and cosmetics generated by applying decision tree association rules.

Therefore, it is more effective to put such an integrated model into practice than simply applying models out of one method.

Acknowledgements

Research on the social science of the Ministry of education, the research on the evolution mechanism of network group events and emergency strategy research under the micro blog platform, No. 13YJCZH124 and No. 222201221102.

References

[1] Silversteinc, B.M. (1997) Beyond Market Basket: Generalizing Association Rules to Correlations. *Proceedings of the* 1997 *ACM SIGMOD International Conference on Management of Data* (*SIGMOD*97), Tucson, Date, 265-276.

[2] Berson, A. and Smith, S. (2001) Building Data Mining Applications for CRM. Posts & Telecom Press, Beijin.

[3] Alfredo, V.A. (2000) Methodology for the Characterization of Business-to-Consumer E-Commerce. John Moores University, Liverpool.

[4] Tsiptsis, K. and Chorianopoulos, A. (2010) Data Mining Techniques in CRM: Inside Customer Segmentation. John Wiley and Sons, Hoboken. http://dx.doi.org/10.1002/9780470685815

[5] Rygielski, C., Wang, J.C. and Yen, D.C. (2002) Data Mining Techniques for Customer Relationship Management. *Technology in Society*, **24**, 483-502. http://dx.doi.org/10.1016/S0160-791X(02)00038-6

[6] Han, J.W. (2002) Data Mining Concepts and Techniques. Machinery Industry Press, Beijing.