International Journal of Hybrid Information Technology
Vol. 9, No.8 (2016), pp. 425-432
http://dx.doi.org/10.14257/ijhit.2016.9.8.38

The Analysis of the Influences of Sales of Merchandise Based on the Decision Tree

Wang Yonggang

School of information science and technology ,Zhengzhou Normal University, China
wyghaha@163.com

Abstract

E-commerce, as a new business model, has become a main sales model now. Exploring the historical data of merchandise sales means a lot to the sales of products . Nowadays there has a new research tendency which is to apply the data mining algorithm to the sales of merchandise . The article herein will use the C4.5 algorithm of decision tree and prime attributes of merchandise to build the model of decision tree, then accomplish the analysis of influences of merchandise sales. The model of merchandise sales can effectively gain the main influences of merchandise and rules of sale through examples and decision tree.

Keywords*: e-commerce ; decision tree ; information gain; influencing factors of sales*

1. Introduction

With the development of Internet and mobile devices, e-commerce, as a new business model, has also boomed and brought huge economic benefits to electric commercial enterprises. The vicissitude of electric commercial enterprises can be seen initially from the operation data. As a result, exploring the valuable data of e-commerce has been an irresistible trend. Taking advantages of data and making strategy based on these data will bring inestimable economic benefits for electric commercial enterprises.

Traditional ways of data analysis can not deal with complicated business data. It is necessary to use the data mining algorithm to extract more valuable information for enterprises, which will give enterprises data support when making strategy. Data mining technology has gradually been applied to commercial operation since recent years , for example, the application of association rule in e-commerce, which could discover the relevance of some products through huge amounts of historical sales data existing in the recommended system and extract the new rules , can provide the basis for the recommendations of subsequent products. [1] Using the time series model to predict the sales volume, which can provide the data basis for the government and enterprises, enables to make strategy or adjust of regulations according to the predicted value [2].

Decision tree algorithm is one of the core algorithm of data mining technology, widely used in classification, prediction and rules extraction. It can discover potential information through the purposeful classification of huge amounts of data. Decision tree classification algorithm, as a common classification algorithm , has been widely used in the fields of finance and medical treatment [3]. ID3 algorithm [5] is one of the classic algorithms, which was come up with by Qninian in 1986 based on information entropy of decision tree algorithm. He introduced the information science into decision tree algorithm and took the information entropy as the standard of selecting test attribute in order to classify the samples and build a decision tree to predict how to divide the all instance memory by test attributes. Decision tree C4.5 algorithm is also come up with by Quilnan, which is the extension of ID3 and the improved version of ID3. C4.5 algorithm increases the solutions

ISSN: 1738-9968 IJHIT

when the continuous attributes and attribute values have vacancies and has a mature method of tree pruning.

The article herein integrating the features of e-commerce business, will help decision makers to comprehend the main factors of merchandise sales, make reasonable production and marketing strategies and predict the sales with rules by the means of discovering the influential factors of merchandise sales, using the C4.5 decision tree classification algorithm to classify the historical data of merchandise sales and building the rules between several influential factors and sales target. The method discussed here may build the classification models in accordance with information entropy in training dataset. When using this method to make classification, it just needs to compare each attribute values of merchandise in a top-down way to finally gain the leaf nodes marking the types of sales status. The results indicate that the application of C4.5 decision tree algorithm used in the influential factors of merchandise sales has a clear advantage on both stability of classification and treatment efficiency of data.

2. Relative Definition

Definition 2.1: E-commerce is mainly about the activities performing through the internet, which is a new commercial operating model and presents in the following ways: Firstly, it is a worldwide commercial business operated through an open internet. Secondly, it is a commercial business operated by the buyers and sellers without meeting one another. It makes customers shop on line, the transactions between merchants, on-line payments, and other commercial activities, such as transactions, financial activities and relative comprehensive services available. [4]

Definition 2.2: The classification of decision tree indicates a rule of classification to predict an expression of decision tree from a group of unordered and ruleless examples. Decision tree classification method uses a top-down recursive fashion, compares the attribute values in the internal node of decision tree and estimates the subordinate embranchments of the node mentioned above according to different attribute values so as to draw conclusions of decision tree leaf nodes.

Definition 2.3: Information Gain

Information gain enables to measure the ability of data samples with more than one attribute division. The more information gain is, the property as a root node of a tree will simplify the tree and use the entropy[3] to measure the uncertain information with the corresponding attribute. Assume that S presents the sample set , what its expected information will mark with entropy, expressed as E(S):

$$E(S) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

the "m" in the expression means the number of different attribute, with presentation Ci(i=1,2,…,m); "pi" means the probability of random groups of samples belonging to "Ci"; Attribute X has different attribute values marked as "n", using the attribute X to divide the sample set to "n" subsets with the expression $E(X) = -\sum_{j=1}^{n} w_j E(S_j)$, "wj" means the weight of subset "sj", $w_j = \frac{|S_j|}{|S|}$, $|S_j|$ means the number of sample subset S_j, $|S|$ means the totality of sample sets.

The information gain get from the embranchment of attribute X expresses as :

$Gain(X) = E(S) - E(X)$

Definition 2.4: the proportion of information gain means the ratio of information gain and segmented information .The computational formula presents below:

$$GainRatio(X) = \frac{Gain(X)}{SplitI(X)}$$

X means the attribute of influential factors during the sales of products, $SplitI(X)$ means segmented information calculated from the concept of entropy,

$$SplitI(X) = -\sum_{j=1}^{n} p_j \log_2(p_j)$$

Definition 2.5: Prime influential factor is the one that could influence the decisions of customers when the sales environment of products is same.

Definition2.6: Prune the Decision Tree

Aiming at the excessive fitting of the sample datum , we propose the pruning of the decision tree. We use the statistical approach to prune the uncertain embranchments for accelerating the recognition capability and classification speed, which substantially indicates the deletion of abnormal samples and disturbances in the training set. After pruning, the smaller the decision tree is, the less complication of it will be, and as a result, it will be easier to understand. The common ways to pruning are pre-pruning and post-pruning. [8]

(1)Pre-pruning

Pre-pruning method is to stop building and prune the tree in advance. Once the building process stops, the node will become leaf node, which may possess the most frequent classification of all the sample subsets. Information gain and GINI index may be utilized in building the decision tree. Moreover, measurements for selection of attribute such as information gain ratio may estimate the good and bad of the production of embranchments. If the value of classification of training set of one segmented node is lower than the preset threshold value, the further division of the subset will cease immediately. Therefore it seems hard to select a reasonable threshold value. The higher threshold value may lead to a decision tree with higher simplification while the lower threshold value may lead to one with lower simplification.

(2)Post-pruning

Post-pruning method is to prune redundant embranchments of a grown tree. It may mark the substitutive subtree whose classification containing the highest quantity of training set through deleting the embranchment in one node and substituting leaf for subtree of the node. The expected classification error rate of subtree in each internal node may be calculated respectively both before and after pruning from the bottom of the tree. If the expected classification error rate rises after pruning, the pruning discussed here should be dropped and the subtree may be saved, otherwise the subtree in the node mentioned may be pruned. An independent data set will be chosen in all the candidate decision trees that have been pruned. Moreover, classification accuracy rate of each candidate decision tree may be estimated. Finally, one decision tree with the lowest expected classification error rate will be chosen as the best decision tree.

Definition 2.7: Evaluation Index of Decision Tree

The complexity and accuracy of decision tree are the most two important factors in the study of decision tree algorithm. The following quantitative evaluation standards are presented by article. [8]

(1)Predictive Accuracy :

The index will mainly describe the ability of classification model to predict new or unknown data. Predictive accuracy is the prime concern of describe makers. The reason why they choose the classification model can be expressed as following: Classification model can operate data and classify at the same time to meet the needs of users in huge amount of data. The differences of the accuracy of information, which is operated by the classification model, will affect on the accuracy of the decisions of decision makers' to a great extent.

(2)Simplicity of Description:

This index is proposed to target at the description of classification model and comprehensibility of the description. The ultimate purpose of classification model is to provide convenience to decision makers, so the more simplified the description of model is for decision makers to use, the more popular it will be. For example, the description of classification model provided by classifier construction using the rule representation seems more simplified and easier to understand while the description results of visual networking are relatively hard to understand. As a result, the further use of that is limited.

(3)Complexity of Calculation :

Complexity of calculation depends on details. In the exploration of data, the operands are the bulk of datum, so the complexity of space and time will be a truly significant link which will directly affect the generation and utility of model and the calculation cost as well.

(4)Power of Model:

The power of model may serve as a supplement of predictive accuracy. When there exist noise and shortage and damage data, it enables to classify the datum accurately. As mentioned above, the operands of data exploration are the bulk of database, and sometimes with the shortage and damage datum, noise datum and also redundant datum in, it is necessary that the established model shall fully adapt to all these complicated situations.

(5)Operation Scale:

Operation scale means the ability of building a model in the huge bulk of datum and achieves the accuracy of classification model. The operands of datum exploration may be plentiful, so it demands that the established model can adapt to the different scales of datum.

3. The Model of Influential Factors of Merchandise Selling Based on Decision Tree C4.5 Algorithm

3.1. Confirm the System of Factors Influencing the Sales of Merchandise

There are many factors that will influence the sales of merchandise, but the effect of each factor is different. When consumers are making decisions, factors with different attribute have various effects on consumers. What we want to do is to find the most influential factor of all, which will promote the sales of merchandise. Here we set the selling of leisure wear as a example, after research we summarize main attribute factor set that may have the greatest impact on the sales of leisure wear represented in chart (1), expressed as $A=\{A_1, A_2, ..., A_n\}$, classification attribute set as sales attribute.

Table 1.The Detailed Statement of Attribute Factors

Influential factors	Description	Value
Price	Price per Unit	Low-end, Midrange, High-end
Fabric	Fabric of Product	Woolen, cotton, terylene, silk, flax,
Style	Style of Product	Luxury, sweet, cartoon, simple, sports, sexy, casual, noble
Pattern	Patterns of Product	Plants and flowers, cartoon elements ,pure color, check pattern , letters, shivering, dots, striae, hearts, geometry ,red star, leopard print
Thickness	Thickness of product	Thicker, ordinary, thin, thinner
Sales volume	Sales Volume of Product	Boom, ordinary, dull of sale

3.2. Establish the Decision Tree of Influential Factors of Merchandise Selling

(1)Calculating the Information Gain Ratio

C4.5 decision tree of influential factors of merchandise selling may choose the highest information gain rate as the test attribute of the current node. According to the Definition 4, we can calculate the information gain rate of each influential factor successively.

Table 2. The Calculations of Information Gain Ratios

Attribute Factors	Information Gain Ratio
Price	0.53
Thickness	0.26
Fabric	0.11
Style	0.08
Pattern	0.03

(2)Establishing Decision Tree Model

The factors, influencing the sales of merchandise, comprise several parts, such as culture, politics, enterprises, economy and characters of merchandise. The article here will mainly discuss the influences of attribute of merchandise on sales, and take the other influences as the same in some extent.

Establishing decision tree model, firstly we will calculate the information gain ratio of each attribute influential factors according to the training data set, then choosing the highest information gain ratio as the best test attribute, which is the decision node of the generation of the current decision tree. We will establish the decision tree by parity of reasoning. The process of the establishment of decision tree is represented in the Figure 1 below.

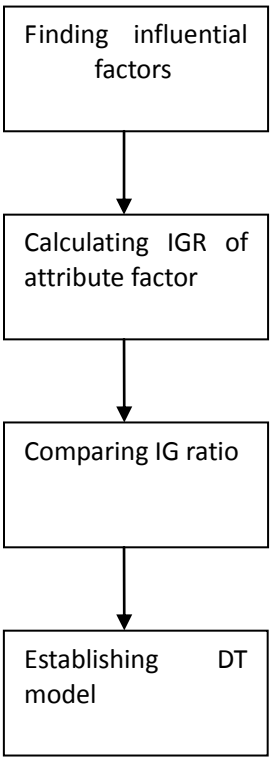

Figure 1. Process of the Establishment of Decision Tree

Process of the Establishment of Decision Tree:

(1)Assuming that decision tree is expressed as $T(B, A, D)$, the "B" in the expression represents sample data set, "A" means the uncertain number of influential factor set, $A = \{A_1, A_2, ..., A_n\}$, and "D" means the field that the influential factor set "A" comes from. D_i means the data range of attribute factor A_i.

(2)Calculating information gain ratio of each attribute factor in terms of Definition 3 and 4.

$$GainRatio(A_i) = \frac{Gain(A_i)}{SplitI(A_i)}$$

(3)Choosing the highest information gain ratio $GainRatio(A_{\max})$ as the goal decision node of decision tree by parity of reasoning.

Firstly, we should arrange the basic datum and take the key attributes which influence the sales of merchandise as the inputs of decision tree model and sales volume as classification goals of decision tree. Secondly, we will gain that "price" influential factor enjoys the highest information gain ratio through the calculation of each information gain ratio of influential factors of sales. As a result, "price" factor will be chosen as the first test attribute to divide the samples. "Price" may treat as root node of decision tree. According to the three attribute values, it may lead to three embranchments, and samples may be divided on the basis of what discussed above. Moreover, we may calculate the influential factor of the highest information gain ratio of sub-sample sets in each embranchment node. Then we may get the decision tree of the influential factor of merchandise's sales by the recurrent method and division of sub-sample sets in terms of the algorithm discussed here represented in the Figure 2 below.

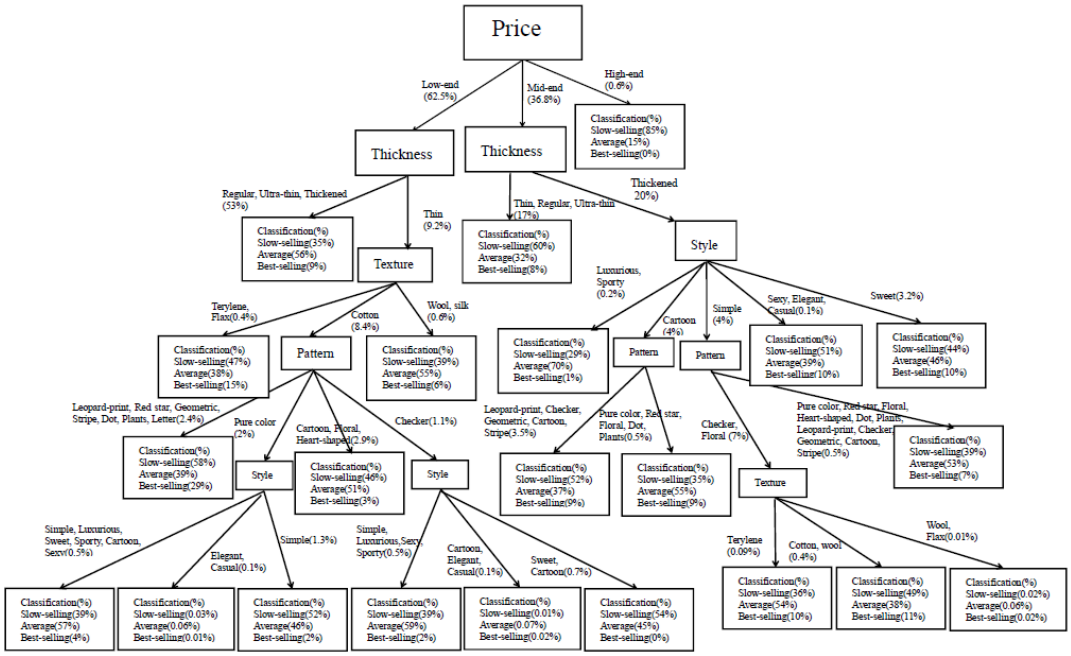

Figure 2. Decision Tree Model of Influential Factors of Merchandise Selling

4. The Analysis of Decision Tree Model

According to the decision tree model in Diagram 2, "price" influential factor is the root node of decision tree, which means it is the main factor that will influence the sales of leisure wear in e-commerce. There are 62.5% of customers buying the low-end leisure wears, and there are 38.6% of customers buying the midrange leisure wears while the proportion of who buy the high-end leisure wears represents only 0.6%, which indicates that people who buy the leisure wears under 200 yuan enjoy the highest proportion while people who buy the leisure wears above 500 yuan enjoy the lowest. As a fact of it, the price of merchandise is the most directive and fundamental factor that will have impact on the sales of leisure wears.

Other factor may have little impact on the sales of leisure wears and may even be ignored. High price will lead to a disappointing sales result, which may account for 85%. Except for the price factor, low-end and midrange leisure wears may also influenced by the thickness of clothes. From the low-end embranchment of decision tree, the thin leisure wears may mainly influenced by the types of fabric, then according to the classification of fabric, the influential factors are patterns and styles. From the midrange embranchment of decision tree, thicker leisure wears may mainly influenced by styles, and the factors of patterns and fabric are in the second place.

Therefore, the main factor of different types of merchandise differs in the process of sales. According to decision tree model, adjusting each attribute of merchandise reasonable may facilitate the sales of products.

5. Conclusion

It becomes more and more complicated and uncontrollable to predict and adjust the sales of merchandise in e-commerce business due to the polytrope of influential factors. The article herein proposes a model of factors that will influence the sales of merchandise based on decision tree. The rules of sales may be extract from the decision tree model of influential factors. Adjusting the selling mode of merchandise according to the extracted rules may increase the benefits. It is significant for enterprises and manufacturers to

analyze the features of datum of merchandise sales and influential factors. Only if we can raise the accuracy of the analysis, enterprises and manufacturers will treat the sales of merchandise more objectively and improve the sales of merchandise in line with the features of merchandise and control of influential factors.

The model discussed here may clearly and correctly represent the main influential factors of merchandise and the rules of sales through the analysis, however, there may exist mistakes when extracting the rules of sales, which are because of incapacity to fully master influential factors of the sales of each product.

Therefore, concerning about the complexity of e-commerce business market, we can integrate decision tree algorithm with other data mining algorithms, such as clustering algorithm, neural network algorithm, which may contribute to the determinacy of influential factors of merchandise sales to some extent. We may make further research later.

References

[1] Y. Zhao and C. Liang, "Recommendation system based on association rules in E-commerce business", Value Engineering, vol. 25, no. 5, **(2006)**, pp. 82-85.
[2] S. Guo, L. Wang and K. Huang, "Research of prediction of automobiles' sales volume based on time serial model", Mechanical Engineer, no. 05, **(2013)**.
[3] X. U. Peng and S. Lin, "Software I O, *et al.* Internet Traffic Classification Using C4.5 Decision Tree", Journal of Software, vol. 20, no. 10, **(2009)**, pp. 407-414.
[4] Z. Guo, "The research of Credit mechanism in e-commerce", Beijing Jiaotong University, **(2012)**.
[5] D. Thakur, N. Markandaiah and D. S. Raj, "Re optimization of ID3 and C4.5 decision tree", International Conference on Computer and Communication Technology, **(2010)**.
[6] J. Han and Y. Gu, "Study on handing rang inputs methods on C4.5 algorithm", Computer Science-Technology and Applications, **(2009)**.
[7] L. D. Safavian Sr, "A survey of decision tree classifier methodology", IEEE Transactions on System, Man and Cybernetics, vol. 22, no. 5 /6, **(1998)**, pp. 660-674.
[8] L. Zhen and W. Wei, "The improvement of data mining algorithm based on decision tree" Southwest Jiaotong University, **(2005)**.

Author

Wang Yonggang, He was born in Henan in the March of 1982,the Master of software engineering of Zhengzhou University, and the PhD in education leadership and management of East China Normal University, he also is a Graduate tutor of the school of information science and technology of Zhengzhou Normal University, and his main research direction is e-commerce and education management.

