Application of ID3 Algorithm in Customer Management about Online Bookstore

Ming-jun Zhang

School of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou 730020, China

zhangmjlz@163.com

Keywords: The classification method of decision tree ; ID3 algorithm ; customer management.

Abstract. ID3 algorithm is one of the most important algorithm of decision tree classification method, Because of its simple, easy to implement, has been widely applied in various fields. Based on introducing the basic principle of ID3 algorithm, this paper has deeply analyzed the practical application of ID3 algorithm in the online bookstore customer management. By this algorithm, we can accurately classify the customer of online bookstore, so as to provide a basis for developing individualized service.

Introduction

ID3 algorithm is one of the most representative algorithms in decision tree classification algorithm. The decision tree classification method is to classify the data through a series of rules. It is to make use of information theory of mutual information (information gain) for finding database with attribute fields in the maximum amount of information, establish of a node of decision tree. Then according to the different values of attribute field to build the tree branch, in each descendent is repeated to establish the process of lower level nodes and branches of a tree.

Due to the practicality of ID3 algorithm, in recent years, many experts and scholars have discussed and researched on the optimization and the application of the algorithm. Such as Chai Hongtao[1] presented that ID3 algorithm on the basis of information resource classification management mapping model is studied in many fields. Han Chengyong[2] ID3 application of attribute reduction, packet count and database technology based on data reduction and storage process of algorithm to optimize and improve the design. Luo Yuzi et al[3] based on the comparison and analysis of ID3 algorithm in order to enhance design. Zhao Yonghui [4, 5] on the improvement of the ID3 algorithm and the application of data mining in the university students' loss of data mining. Du Liying [6] starts with the basic theory of ID3 algorithm. The specific application of the algorithm in the bank's individual customer management is analyzed in detail. Visible, the current research in this area has a lot of results. This paper will discuss and analyze the application of ID3 algorithm in customer management about online bookstore.

The basic theory of ID3 algorithm

ID3 algorithm

The ID3 algorithm was proposed by Quinlan in 1986. They used information entropy theory. The attribute value of the maximum information gain of the current sample set is selected as the test . The partition of the sample set is based on the value of the test attribute. How many different values of the test attributes will be divided into the number of subsets of samples. At the same time, the decision tree is corresponding to the node of the sample set to grow a new leaf node. The algorithm will be listed as follows[7] :

Algorithm (Generate_decision_tree)

(1) Create a root node N;

(2) If all the samples in this node belong to the same class C, then

(3) Return N as a leaf node, marked as class C;

(4) If attribute_list is empty, then:

(5) Return N as a leaf node, and mark the most of the categories in which the node contains the largest number of categories;

(6) Select one of the most information gain properties from test_attribute attribute_list;

(7) And the node N is marked as test_attribute

(8) For each of the known values of a_i in test_attribute, prepare for dividing the sample set containing node N;

(9) According to the condition test_attribute $= a_i$, a corresponding branch is generated from the node N to indicate the test condition;

(10) A set of samples obtained by s_i for conditional test_attribute $= a_i$;

(11) If the s_i is empty, the corresponding leaf node is labeled as the most of the categories in the sample contained in node;

(12)Otherwise, the corresponding leaf node is marked as the return value of Generate_decision_tree.

Related definitions

Define 1. information Let *S* as a set of containing *s* data samples. Class properties can take *m* different values, corresponding to *m* different categories $C_i, i \in \{1,2,\cdots,m\}$. Suppose s_i is the number of samples in class C_i. The amount of information needed to classify a given data object is:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log(p_i) \quad (1)$$

where p_i is the probability denote as an arbitrary data object belongs to C_i, can be calculated by s_i / s, while the *log* function is base 2, because the information in the information theory are by-bit encoding.

Let a attribute *A* takes different values *v* namely $\{a_1, a_2, \cdots, a_m\}$.By using the attribute *A*, we can divided the set *S* into *v* subsets $\{S_1, S_2, \cdots, S_v\}$, where S_j containing data sample *S* of attribute *A* with the values a_j. If the attribute *A* was selected as test, let s_{ij} denote as the number of sample of subsets S_j belong to class C_i. Then take advantage of attribute *A* divided the current sample set for wanting information. It was calculated as

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj}) \quad (2)$$

where $(s_{ij} + \cdots + s_{mj}) / s$ was denoted as the weight of *j* subsets, which was made up of centrally by all sub-attributes *A* number of samples taken of a_j value divided by the total number of samples set *S*. The result of E(A) is smaller, it means that subset is better result. For a given subset of S_j, the amount of information it is:

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log(p_{ij}) \quad (3)$$

where $p_{ij} = s_{ij} / |S_j|$, that is the probability of a data sample of subset S_j belongs to class C_i.

Define 2. information gain The current branch node corresponding gain sample set partitioning information acquired using the attribute A is:

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \quad (4)$$

Obviously, Gain(A) is considered to be a set of samples were divided according to the value of attribute *A* decrease in entropy obtained. ID3 algorithm to calculate information gain of each attribute, and pick out the greatest attribute of information gain as a given set *S* test attribute and thereby

generate a corresponding sub-node. Generated node is marked as the corresponding attribute, and generates a corresponding branch respectively according to different values of the properties, each branch represents a subset of the sample is divided.

The application of ID3 algorithm in cusomer management about online

With the development and improment of internet technology , e-commerce quickly penetrate into all-round of our life, in all areas, for example book sales. Online sales of books is not only to develop and complement traditional sales channels, but also the future direction of book sales. Domestic online sales of books started in the 20th century, the mid-1990s. In 1995, mainland China's first online bookstore - Chinese bookstore officially launched operations; In early 1997, Hangzhou Xinhua Bookstore move the traditional sales on the Internet, open up the country;In 1999, Beijing Book online bookstore building officially opened, and so on. From the end of 1999 to the first half of 2000, the domestic Internet companies, venture capital firms set off a wave of the founder of the online bookstore. Online book sales across the country to get involved in the Beijing Xidan Shopping Mall, Dangdang, Beijing Book Building, Shanghai Bookstore, Guangzhou Book Center. Online bookstore in the course of business, how to improve the relationship with customers, and maximize customer value, it would be the operators who need to solve the problem. We will ID3 algorithm is applied to the online bookstore customer management, decision tree methods to construct model for customer value analysis, to find the most valuable customers, and thus carry out targeted promotional activities to provide a more personalized service.

ID3 algorithm based on a given set of data rows or data object (whose class attribute is known) to construct a decision tree, and then use the properties of the unknown category tree can be classified. Below a certain online bookstore, for example, a detailed analysis of ID3 algorithm application in the customer management.

Construct training set

According to user information and logs certain online bookstore database, extract customer activity information part of building a data set (see Table 1):

Table 1 Certain online bookstore book sales information

Client	Amount of consumption	Book Information	Students	Book species	Buy
1	＜50	Group	Yes	Science and Engineering	Yes
2	≥100	Group	No	Social Sciences	No
3	50～100	Promotions	Yes	Magazines	Yes
4	50～100	Ordinary	Yes	Science and Engineering	No
5	＜50	Promotions	No	Magazines	Yes
6	＜50	Ordinary	No	Social Sciences	Yes
7	≥100	Group	Yes	Magazines	Yes
8	≥100	Ordinary	Yes	Science and Engineering	No
9	＜50	Ordinary	No	Science and Engineering	Yes
10	50～100	Promotions	No	Social Sciences	Yes

Computing information gain

The application of ID3 algorithm generates the execution process of a decision tree algorithm is described as follows:

(1) Category attribute of the training set "whether buy" has two different values, i.e. {yes, no}, so there are two different classes ($m=2$). Suppose class C_1 corresponds to "Yes", and C_2 corresponding to "NO", there are seven sample in class C_1, class C_2 has three samples. according to the Eq.1 to calculate the information needed to classify a given sample:

$$I(s_1,s_2)=I(7,3)=-\frac{7}{10}\times\log_2\frac{7}{10}-\frac{3}{10}\times\log_2\frac{3}{10}=0.881$$

(2)Calculate information for each attribute (entropy). From the attribute"the amount of consumption" began, according to the attributes of each value in "yes" category and the distribution of "no" category, can be calculated for each distribution information corresponding to:

For the "amount of consumption "< “50”： $s_{11}=4$, $s_{21}=0$

$$I(s_{11},s_{21})=I(4,0)=-1\times\log_2 1=0$$

For the "amount of consumption " =“50～100”： $s_{12}=2$, $s_{22}=1$

$$I(s_{12},s_{22})=I(2,1)=-\frac{2}{3}\times\log_2\frac{2}{3}-\frac{1}{3}\times\log_2\frac{1}{3}=0.918$$

For the "amount of consumption" ≥ “100”： $s_{13}=1$, $s_{23}=2$

$$I(s_{13},s_{23})=I(1,2)=-\frac{1}{3}\times\log_2\frac{1}{3}-\frac{2}{3}\times\log_2\frac{2}{3}=0.918$$

Then using Eq.2 can be calculated only if based on the attribute "amount of consumption" of the sample to be divided, obtained for a data object classification required information entropy:

E(amount of information)=4I(s_{11},s_{21})/10+3I(s_{12},s_{22})/10+3I(s_{13},s_{23})/10=0.551

(3)Computing information gain. According to the Eq.4 to calculate the use of property " amount of consumption " the collection of samples for classification obtained information gain:

Gain(amount of information)=*I*(s_1,s_2)-*E*(amount of information)=0.881-0.551=0.33

Similarly, calculate the "textbook approach", "whether the student", " book species " information gain:

Gain(Book Information)=0.207 *Gain*(Student)=0.034 *Gain*(Book species)=0.206

ID3 algorithm uses the "information gain" parameter to evaluate the significance of attributes. The use of a node tree structure of the property, and test the properties they represent all the value in the node, in order to obtain the various branches of the node, these branches of the original data set is divided into several sub-data set . If the data row of a node contained herein are the same category, then the leaf node of the decision tree and the node is marked as the appropriate category.

Construction of decision alternatives

According to the calculation results, the properties "amount of consumption" obtained maximum gain. So the property will produce current branch node as a test property. Generates three different branches according to the attribute "amount of consumption" of different values. The current sample collection is divided into three distinct subsets. As shown in Fig.1.

Seen from Fig.1, when the "amount of consumption" <50 of the sample category 50 are "Yes" category, resulting in a leaf node at the end of the branch and marked "yes". According to the training sample set as shown in Table 1, eventually produce the decision tree shown in Figure 2.

By analyzing the ID3 algorithm, we can see "the amount of consumption" is the most important factor in the decision tree branches, followed by "textbook fashion", " book species ", "whether the student" and so on. According to the decision tree shown in Fig.2, the following conclusions:

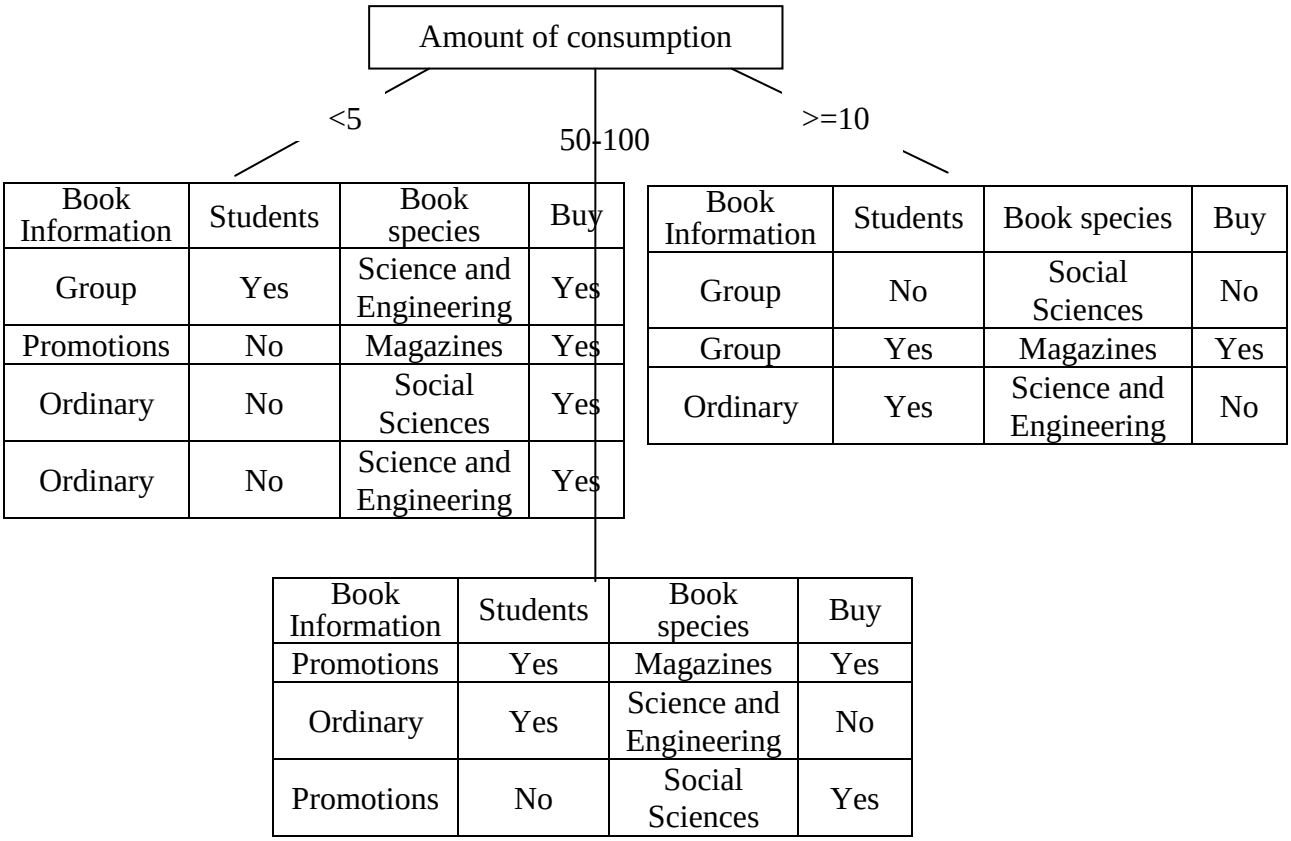

Figure 1 Schematic diagram of attribute selection "consumption amount" of the corresponding branches

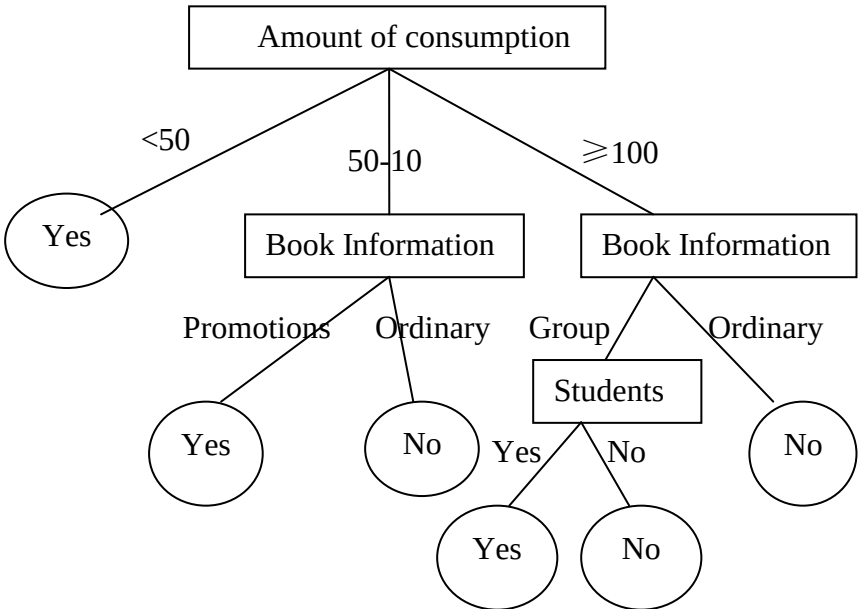

Figure 2 The resulting decision tree

(1)Easily lost customers: the amount of consumption between 50 to 100 customers in the normal way using the textbook way; the consumption amount is greater than or equal to 100, to participate in buy, not a student customers; consumption amount is greater than or equal to 100, the ordinary way purchases client.

(2) Loyal customers: the amount of consumption of less than 50 clients; the amount of consumption between 50 and 100, participation in promotions of customers; consumption amount is greater than or equal to 100, to participate in buy, is the student's customers.

According to the analysis result, the loss of customers to easily launch targeted offers, promotions and other activities, so as to attract customers, customer retention purposes; For loyal customers, their characteristics can be further analyzed to provide better personalized service.

Conclusion

ID3 algorithm is a classic decision tree classification algorithm, decision tree is most commonly used to generate specific method. The algorithm is applied online bookstore customer management, you can dig out a lot of potential, implied, useful customer information. On this basis, build predictive models for different customers to classify, identify the causes of the loss of customers, the implementation of targeted strategies to improve customer loyalty.

Acknowledgements

This work was financially supported by Lanzhou University of Finance and Economics Teaching and Research Project(No.20140201).

References

[1] Chai Hong-tao,Li Jian-hua,Shen Di: Study on information resources classification management mapping model based on ID3 algorithm. *Computer Engineering and Design*. 2013(3): 1082-1086.

[2] Han Cheng-yong: The design to improvement of ID3 algorithm based on data reduction and stored procedure. *Natural Sciences Journal of HaRBin Normal University.* 2013(4):52-54.

[3] Luo Yu-zi ,Fu Xing-hong:Data mining ID3 decision tree classification algorithm and its improved algorithm. *Computer Systems & Application*. 2013(10):136-138.

[4] Zhao Yong-hui: Research on datd mining in loss of college students based on fuzzy ID3 algorithm. *Computer Era*. 2014(3):36-38.

[5] Zhao Yong-hui:Improvement research of data mining algorithm ID3. *Computer Development & Applications*. 2013(4):61-63.

[6] Du Li-ying: Application of decision tree ID3 algorithms on bank customer relationship management. *Journal of JiLin Institute of Architecture & Civil Engineering.* 2013(12):49-51.

[7] Zhu Ming, in: Data Mining. *Press of University of Science and Technology of China*.2002.