ISSN 2616-5775 Vol. 5, Issue 3: 101-106, DOI: 10.25236/AJCIS.2022.050315

Research on Product Recommendation Based on Association Rules and Improved Apriori Algorithm

Jiayi wang[1,a], Mengjia Jiang[2]

[1]School of Labor Economics, Capital University of Economics and Business, Beijing, 100070, China
[2]School of Maths and Physics, Chengdu University of Technology, Chengdu, 610059, China
[a]ebaodou@126.com

Abstract: *With the rapid development of the Internet era, online sales services have gradually become an indispensable part of people's daily life. In the face of huge network data, how to find products that can maximize the utility of customers and form product recommendations that are beneficial to merchants has become an important issue. In this paper, through the in-depth analysis of the commodities and customer purchase data of a bicycle commodity store, the Apriori algorithm is used to mine the association rules of the data, and the commodity combination with strong correlation is obtained. At the same time, the traditional Apriori algorithm has certain limitations, that is, the algorithm only considers the probability of the transaction, but does not consider the weight of different products, and then we introduce the intuitionistic fuzzy number to represent the weight of the item. Improvements have been made, resulting in a more accurate and effective product recommendation combination. The implementation of this algorithm can be widely used in the commercial field to achieve the purpose of using best-selling commodities to drive relatively non-selling commodities, and at the same time maximize the utility of consumers, which also brings greater benefits to businesses, thus forming a healthy network. The sales structure has important practical significance to the current era.*

Keywords: *Product recommendation; Apriori algorithm improvement; Consumer behavior; Fuzzy numbers; E-commerce*

1. Introduction

In the current Internet + era, more and more network services have entered our lives, of which e-commerce is a more obvious part. Compared with traditional sales methods, online sales can make it easier for customers to contact more kinds of goods and have more choices. As an e-commerce merchant, the most important thing is to provide customers with the most matching product recommendations that customers are most interested in. In this paper, by analyzing the consumption behavior and consumption preferences of customers, combining association rules and Apriori algorithm, and introducing intuitionistic fuzzy numbers to represent the weight of items, the Apriori algorithm is improved, and the problem of the limitations of traditional Apriori algorithm in mining is solved. A more accurate product recommendation method is developed, so that e-commerce can better serve customers, maximize customer utility, and at the same time greatly increase the profits of enterprises, forming a better network sales structure.

Previously, Lin Sui [1] used weights to distinguish the importance of customer types, and then dug out the association rules between customer types and customer consumption amounts, which better help the managers of related websites to rationally schedule manpower and optimize staffing. Yu Yan [2] has proved the effectiveness of the effective time probabilistic association rule algorithm through experiments, and applied it to product recommendation. The method of sorting the candidate set according to the minimum support of item attributes is adopted to reduce the overhead of algorithm operation. and applied it to product recommendation. Jia Hang [3] proposed an algorithm based on the similarity of user consumption behavior to recommend products; Zeng Lei [4] proposed a Maping-Apriori algorithm model based on mapping. In summary, all the methods mentioned above are avoiding the drawbacks of the Apriori algorithm, providing more accurate product recommendations, and better improving work efficiency in real life.

2. Data processing

This article uses the product data of a bicycle shop from the GitHub open source database. By previewing the data, we can see that the data content includes the product name, the name of the purchaser, and the purchase order. By using Excel to sort and sort the data, get the order in which the same customer buys different products.

Data preprocessing: Due to the huge amount of data, we first use python to perform duplicate value detection to determine a total of 37 commodity types and their specific names in this set of data, and then sort and compare the repetitions of the same commodity name. The analysis shows that the top 15 best-selling items in this bicycle shop are: 'Sports Helmets', 'Sports Bottles', 'Repair Tools', 'Mountain Bike Tubes', 'Mountain Bikers', 'Road Bike Tubes', 'Bicycle Cap', 'Mountain Bike Fender', 'Mountain Bike Bottle Cage', 'Road Bike Bottle Cage', 'Long Sleeve Cycling Jacket', 'Short Sleeve Classic Bike Jacket', 'Racing Road Bike', 'Touring car inner tube', 'Half palm gloves'. From this, the types of goods that customers prefer to buy are obtained.

3. Model Establishment

3.1. Initial analysis

After preliminary analysis, within this shopping area,top 15 best-selling recommended items that are most popular with customers in this bicycle shop are as follows:

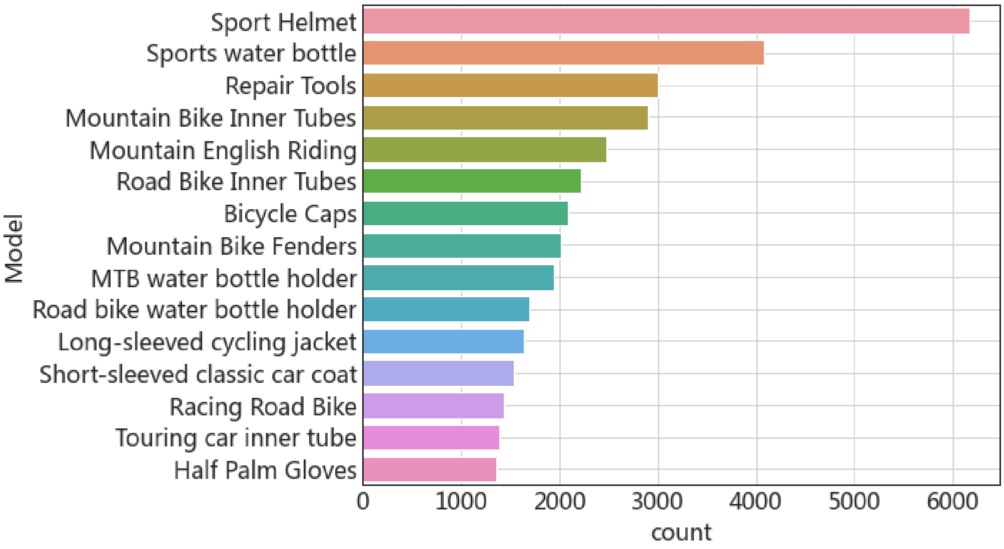


Figure 1: The top 15 best-selling recommended items

3.2. Model introduction

"Apriori algorithm is a commonly used algorithm for mining data association rules. It is used to find data sets that frequently appear in data values, and finding the patterns of these sets can help us make some decisions. Apriori algorithm is widely used, it can be used for price analysis of the consumer market and predict the consumption habits of customers. The core idea of the Apriori algorithm is to use a layer-by-layer iterative search method to mine the k-order frequent itemsets until the highest-order frequent itemsets are found. Then, the association rule mining is carried out through the obtained frequent itemsets, so as to successfully mine the relationship between the target data [4] The mining steps mainly include the following two steps. Firstly, find all frequent itemsets according to the support degree. Secondly, generate association rules according to the confidence degree.[5]

The specific operation method is:

First, find the set of frequent "1 itemsets", and denote this set as L1. Then use L1 to find the set L2 of frequent "2 itemsets", and L2 is used to find L3. By analogy, it is known that the "K item set" cannot be found, that is, a larger frequency set cannot be found, and a database scan needs to be performed every time an LK is found.

Then, according to the definition of confidence, the following association rules are generated:

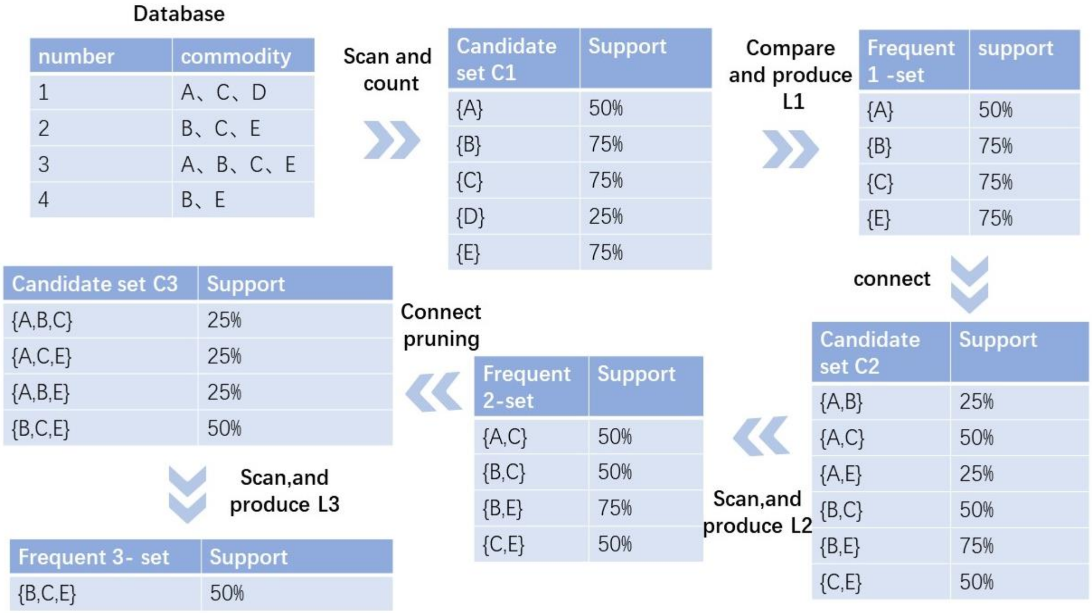
(1) For each frequent itemset L, generate all non-empty subsets of L;

(2) For each non-empty subset S of L, if

$$P(L)/P(S) \geqq min\_conf,$$

Then the output rule "S àL-S".

In order to better illustrate the apriori algorithm, the following examples and flowcharts are used to illustrate the description. Assuming that there is a database as shown in the figure, its mining method according to the Apriori algorithm is as follows:



number	commodity
1	A、C、D
2	B、C、E
3	A、B、C、E
4	B、E

Candidate set C1	Support
{A}	50%
{B}	75%
{C}	75%
{D}	25%
{E}	75%

Frequent 1 -set	support
{A}	50%
{B}	75%
{C}	75%
{E}	75%

Candidate set C2	Support
{A,B}	25%
{A,C}	50%
{A,E}	25%
{B,C}	50%
{B,E}	75%
{C,E}	50%

Frequent 2-set	Support
{A,C}	50%
{B,C}	50%
{B,E}	75%
{C,E}	50%

Candidate set C3	Support
{A,B,C}	25%
{A,C,E}	25%
{A,B,E}	25%
{B,C,E}	50%

Frequent 3- set	Support
{B,C,E}	50%

Figure 2: The mining method according to the Apriori algorithm

3.3. Model establishment

The shopping basket is generated based on the Apriori algorithm, and all the products purchased by the same customer are put into the shopping basket, and the number of shopping baskets is exactly equal to the number of customers in the data set. We import the data and use the dataconvert function to analyze it, and we can see that the data contains 21,255 shopping baskets, that is, 21,255 customers for transactions. According to the permutation and combination, it can be seen that these transactions will generate C_{21225}^{2}an association rule, from which the data that does not meet the support degree or the support degree is too small, can obtain the association results shown in the following table.

Support represents the probability that the set of items {X, Y} will occur in the total set of items. The formula is:

$$Support(X \rightarrow Y) = P(X,Y) / P(I) = P(X \cup Y) / P(I) = num(XUY) / num(I) \quad (1)$$

where I is the total set of transactions. num() is the number of occurrences of a particular set of items in the set of transactions.

Lift represents the ratio of the probability of a condition containing X that also contains Y, to the probability of a condition not containing X but containing Y.

$$Lift(X \rightarrow Y) = P(Y|X) / P(Y) \quad (2)$$

Confidence represents the probability of introducing Y from the association rule "X → Y" if the precondition X occurs. That is, the probability of containing Y in the set of items containing X. The formula is:

$$Confidence(X \rightarrow Y) = P(Y|X) = P(X,Y) / P(X) = P(XUY) / P(X) \quad (3)$$

Table 1: Association Rule Results

lhs		rhs	support	lift	confidence
(ll Mountain Tires)	==>	(Mountain Bike Inner Tubes)	0.021077	4.098245	0.560701
(Road Bike 350)	==>	(Sport Helmet)	0.014679	1.156763	0.335845
(Car wash spray)	==>	(Repair Tools)	0.012468	2.165842	0.306713
(ml mountain outer tire)	==>	(Repair Tools)	0.015055	2.086489	0.295476
(ml road outer tire)	==>	(Repair Tools)	0.012138	2.04703	0.289888
(Touring car inner tube)	==>	(Repair Tools)	0.017031	1.829813	0.259127
(Touring bike outer tire (universal))	==>	(Repair Tools)	0.010115	1.723285	0.244041
(Mountain Bike Inner Tubes)	==>	(ll Mountain Tires)	0.021077	4.098245	0.154058
(Repair Tools)	==>	(Touring car inner tube)	0.017031	1.829813	0.120266
(Repair Tools)	==>	(ml mountain outer tire)	0.015055	2.086489	0.106312

Taking the first row as an example, the first column represents the product purchased by the user - 11 mountain tires; the second column represents another product recommended according to the user's purchase of a certain product - mountain bike inner tube, confidence represents the user The probability of buying both at the same time, lift said that if a user who has purchased 11 mountain tires is recommended to buy a mountain bike inner tube, the probability of the user buying a mountain bike inner tube is about 400% of the customer's natural purchase of a mountain bike inner tube.

Since the Apriori algorithm only considers the probability of occurrence of transactions and does not consider that different items and each item in a transaction has different importance, the mining has some limitations. Moreover, in practical applications, it is more appropriate to use intuitionistic fuzzy number for the weight of items, which not only can better simulate life, but also introduces the concept of hesitation, which can help improve the calculation method of confidence level.

Based on this, the Apriori algorithm is improved by introducing intuitionistic fuzzy numbers to represent the weights of items.

4. Improved Apriori algorithm based on intuitionistic fuzzy

4.1. Algorithm improvement idea

The item weights are set as intuitionistic fuzzy numbers, and the weighted support degree is obtained by applying the formula based on the item weights and the support number of items and introducing the concept of likelihood, and pruning according to the likelihood degree.

Let A and B be any two intuitionistic fuzzy sets, X be a non-empty set, $\mu_A(x)$ be the subordination of the elements in X belonging to A, $v_A(x)$ be the non-subordination of the elements in X belonging to A, and satisfy the condition that: $0 \le \mu_A(x)+v_A(x) \le 1, \forall x \in X$, the mean of the m intuitionistic fuzzy numbers is denoted as the intuitionistic fuzzy number and is denoted as:

$$\overline{A} = \left\langle \frac{\mu_{A_1} + \mu_{A_2} + \cdots \mu_{A_m}}{m}, \frac{v_{A_1} + v_{A_2} + \cdots v_{A_m}}{m} \right\rangle$$

By introducing the possibility degree, the improved Apriori algorithm changes the strategy of pruning only according to the support degree, and prunes the items with the possibility degree less than 0, and prunes the items with the possibility degree greater than 0 and less than 1 according to the confidence degree, so the algorithm takes into account the importance and reliability of the rules, and according to the experiment The reasonableness of the algorithm is proved.

4.2. Algorithm steps

Step 1: Give the minimum support and the weights of m items with intuitionistic fuzzy numbers;

Step 2: Scan the database, calculate the support number of each item, calculate the weighted support and likelihood of the k-item set, delete the k-item set with likelihood less than 0.5 to get the candidate item set, calculate the certainty for the item set with likelihood greater than 0.5 and less than 1, and delete the item set with certainty less than 50% to get the frequent 1-item set;

Step 3: Connect the frequent 1-item set itself to get the candidate 2-item set, calculate the weighted support and likelihood of each item set, delete the items with likelihood less than 0.5, for the items with likelihood greater than 0.5 and less than 1, then calculate the certainty, delete the items with certainty less than 0.5, and get the frequent 2-item set;

Step 4: Execute the join step and branching step cyclically until the resulting item set is empty, the algorithm stops, outputs those frequent item sets, and generates association rules based on the frequent item sets;

Step 5: Calculate the confidence level of each association rule, set the minimum confidence level as 0.5, then compare with the minimum confidence level to get the strong association rule.

4.3. Model Results

Due to space limitation, only some of the strong association rules for complementary items are listed here.

Table 2: Strong association rules

lhs		rhs	support	lift	confidence
(Road bike water bottle holder)	==>	(Sports water bottle)	0.071183	4.635604	0.888954
(Touring bike outer tire (universal))	==>	(Touring car inner tube)	0.035662	13.090553	0.860386
(MTB water bottle holder)	==>	(Sports water bottle)	0.076359	4.360336	0.836167
(hl mountain outer tire)	==>	(Mountain Bike Inner Tubes)	0.043049	5.024695	0.687453
(hl road outer tire)	==>	(Road Bike Inner Tubes)	0.02597	6.585282	0.686567
(ml mountain outer tire)	==>	(Mountain Bike Inner Tubes)	0.034204	4.90651	0.671283
(ml road outer tire)	==>	(Road Bike Inner Tubes)	0.027288	6.25071	0.651685
(Touring car inner tube)	==>	(Touring bike outer tire (universal))	0.035662	13.090553	0.542591
(ll road bike outer tire)	==>	(Road Bike Inner Tubes)	0.024277	5.050274	0.526531

The strong correlation of the improved algorithm is effectively improved and the runtime is reduced by 18.8%. In this way, a more accurate relationship between commodities can be obtained, and by applying it to the commodity recommendation of e-commerce, the purpose of driving relatively non-selling commodities through best-selling commodities can be realized.

5. Conclusion

In this paper, an improved Apriori algorithm is proposed to set the item attributes as intuitionistic fuzzy numbers, so that the information of affiliation, non-affiliation and hesitation are expressed simultaneously, and the candidate item set is obtained by connecting itself, and the weighted support is obtained by multiplying the support number of each item with the weight, and compared with the minimum support, and pruned by combining the possibility and hesitation to obtain the frequent The algorithm is completed by comparing the confidence level of the resulting rules with the confidence level of the rules. Then the strong association rules are obtained by comparing the confidence of the obtained rules. The example shows that the improved Apriori algorithm has some advantages over the classical Apriori algorithm in that it can mine more valuable and meaningful strong association rules, and the item weights are represented in a way that is closer to the characteristics of the merchandising data, and it is easier to derive the weight values of each item.

Biographical notes

Jiayi Wang is an undergraduate student of Capital University of Economics and Business, China. Her research interest focuses on Business Analysis. Email: ebaodou@126.com

Mengjia Jiang is an undergraduate student of Chengdu University of Technology, China. Her research interest focuses on Applied Mathematics. Email: jiangmengjia2022@126.com.

References

[1] Lin Sui, Zheng Zhihao. Research on customer behavior modeling and product recommendation based on association rules [J]. Journal of Guangdong University of Technology, 2018, v.35; No.134 (03): 90-

94.

[2] Yu Yan. Application of Effective Time Probabilistic Association Rules in Commodity Recommendation System [J]. Journal of Nanjing Institute of Technology (Natural Science Edition), 2009, v.7; No.25(01): 58-62.

[3] Jia Hang. Research on Commodity Recommendation Model Based on Similarity Division and Association Rules [D]. Dalian Maritime University, 2020. DOI: 10.26989/d.cnki.gdlhu.2020.000139.

[4] Zeng Lei. Research on Apriori Algorithm in Association Rule Mining [D]. Chongqing Jiaotong University, 2016.

[5] Wang Wei. Research and Improvement of Apriori Algorithm in Association Rules [D]. Ocean University of China, 2012.