American Journal of Mathematical and Computer Modelling
2020; 5(2): 47-50
http://www.sciencepublishinggroup.com/j/ajmcm
doi: 10.11648/j.ajmcm.20200502.13
ISSN: 2578-8272 (Print); ISSN: 2578-8280 (Online)

Parameter Selection Strategy for Frequent Itemsets in Association Analysis

Yuan Hai Yan

Huashang College, Guangdong University of Finance & Economics, Guangzhou, China

Email address:
yan85028@163.com

To cite this article:
Yuan Hai Yan. Parameter Selection Strategy for Frequent Itemsets in Association Analysis. *American Journal of Mathematical and Computer Modelling*. Vol. 5, No. 2, 2020, pp. 47-50. doi: 10.11648/j.ajmcm.20200502.13

Received: March 16, 2020; **Accepted**: April 8, 2020; **Published**: May 12, 2020

Abstract: In data mining, association analysis mainly deals with different associations between things. Different degrees of correlation are usually treated differently in performance. In a production society, people are more interested in understanding the strong relationships between things, while ignoring weaker relationships, thereby making meaningful and valuable decisions. However, people must face several problems. For example, how to use parameters to define strong correlation; how to define meaningful parameters, this article uses experiments to explain the main factors affecting the parameters and how to select parameter values. Find the balance point where the application association produces economic value, then this balance point is a more meaningful parameter. The purpose of this article is to find the support and credibility based on association analysis through dichotomy, and compare the application analysis of the same metric value in different scenarios. Experimental results show that selecting the same parameter value in different scenarios' associated demand analysis (such as attribute association analysis) will not produce the same benefit. In the same scenario, the dichotomy method can make the parameter value close to a more meaningful value. Therefore, how to define the parameters of frequent itemsets to produce the maximum benefit is the significance of this article.

Keywords: Frequent Itemsets, Support, Credibility, Parameter Settings

1. Introduction

In all walks of life, people want to understand the strong correlation between things. By understanding and mastering the correlation, they know more clearly how to make a meaningful and valuable decision. Important content-association analysis. For example, in the telecommunications industry, how to launch a package model suitable for different types of people, so that everyone creates greater value for telecommunications. In this problem, the problems that need to be solved are: what amount of people's telecommunications consumption data can be called a class to define people's consumption habits; how to set package consumption patterns through consumption habits. These questions reflect two important issues in association analysis. First: how to define the frequency of things through support; second: how to set credibility values to weigh valuable links between things. Both of these problems are related to the parameters given [1].

The significance of this paper is mainly how to define the parameters of frequent itemsets to produce the maximum economic benefits. Considering the setting of the support value, it should be considered how large the data scale is to define the support value [2]. In addition to data factors, other factors will also affect the support value of things, for example, the situation of goods sold in different seasons is different. Then the support value setting should be different in different seasons. Therefore, the design support value must have certain principles.

2. Problem Formulation and Solution

There are two more important issues in association analysis that need to be addressed. First: how to define the frequency of things through support; second: how to set credibility values data scale and significance of support value. In the case of a relatively small scale, the occurrence to weigh valuable links between things. For the first problem, we can decompose the problem of of the event has great contingency and uncertainty, so it is important to consider the significance of the data on how large the scale is to stabilize the event. Not all event support values are meaningful. Meaningful means that the application value is higher than the decision cost and can produce economic benefits. For the second problem, the question of the significance of data scale and credibility value is also needed. To explain and answer the above questions, some specific principles must be made. The determination of the size is mainly based on the number of transactions and the transaction product data. The generation of frequent itemsets is determined by the value of support. In association analysis, the frequency of a certain product or product set directly determines whether it is possible to generate inter-products. Association rules, so the determination of support value is carried out experimentally in this article (see Experiment 1 for details). Once the frequent itemsets are generated, they need to be used to generate association analysis. Therefore, the parameter value of reliability in this paper is also explained through experiments (see Experiment 2 for details) [3-6].

Experiment 1: Support Value Experiment.

PURPOSE OF THE EXPERIMENT: Find reasonable support value and filter out meaningful frequent itemsets.

EXPERIMENT CONDITIONS: Data size required for the million level, the dichotomy to find a reasonable support value.

END OF EXPERIMENT: The support value stability factor is less than or equal to the given value.

Experimental Algorithm: Binary Search

Suppose a1=0 and b1=1.

The initial support value sup=(a1 + b1) / 2, n=0 (represents the number of cycles). (1)

Find the support value of an item or item set.

Find the frequent itemsets in the database.

Determine whether the economic benefit index of frequent itemsets is less than or equal to 1.

If it is: a1=sup, n=n + 1, go to step.

Otherwise: b1=sup, n=n + 1.

Is the stability factor for calculating the sup value less than or equal to the given value L?

If yes: End.

Otherwise: go to step.

Experiment 2: Experimental item of credibility value

Experimental purpose: To find reasonable credibility values and to screen meaningful association rules.

Experimental conditions: The data scale is required to be in the level of one million, and a reasonable credibility value is sought by the dichotomy.

Experiment end principle: Find all meaningful association rules.

Experimental algorithm: binary search.

Suppose a2=0 and b2=1.

The initial confidence value con=(a2 + b2) / 2, n=0 (representing the number of cycles). (2)

Read the frequent itemsets according to the support value in Experiment 1.

Find association rules in frequent itemsets

Determine whether the economic benefit index of frequent itemsets is less than or equal to 2.

If it is: a1=con, n=n + 1, go to step.

Otherwise: b1=con, n=n + 1,

Calculate whether the con stability factor is less than or equal to the given value L?

If yes: End.

Otherwise: Go to step.

3. Related Theories and Hypothetical Conditions

For better explanation, the relevant definitions are as follows:

Complete set data: the set containing all itemsets, set to I;

Itemset data: It is a collection of items recorded each time, which is a subset of I. The number of itemset categories is 2 ^ N-1, where N is the number of complete set data [7].

K itemset: The itemset data contains K items, which is called K itemset.

Support: The support of the K-item set is the percentage of the total number of records in the set, recorded as support (X).

Credibility: the degree of dependence of a commodity (set) on another commodity (set) in a K frequent item set [8].

Frequent itemsets: The support of the K itemsets is greater than or equal to the given minimum support (threshold).

P value (economic benefit index)=expected economic benefit / decision cost.

L value (parameter value stability factor)=variance of a1 continuous value

Property 1: All non-empty subsets of a frequent itemset are frequent.

Apriori tailoring rule: If there are certain itemsets that are infrequent, any superset of these itemsets is infrequent,

Data collection strategy: number of commodities N=10 (20), commodity transaction data NSum=100W (200W). The software generates simulation data according to certain rules.

4. Experimental Results and Analysis

4.1. Experimental Results

Table 1. *Parameters of Experiment 1 when N=10 NSum=100W (L value is 0.01).*

Number of experiments	n cycle value	M1 value (number of frequent items)	a1 value (support)	P value (indicator of economic benefits)
1	13	349	0.6377	1.307
2	24	145	0.8354	2.312

Table 2. *Parameters of Experiment 1 when N=20 NSum=200W (L value is 0.01).*

Number of experiments	n cycle value	M1 value (number of frequent items)	a1 value (support)	P value (indicator of economic benefits)
1	16	674	0.6987	1.402
2	28	134	0.9023	3.200

Table 3. *Parameters of experiment two when N=10 NSum=100W (L value is 0.01).*

Number of experiments	n cycle value	M1 value (number of frequent items)	a1 value (support)	P value (indicator of economic benefits)
1	14	121	0.8094	1.307
2	21	57	0.9233	2.312

Table 4. *Parameters of Experiment 2 when N=20 NSum=200W (L value is 0.01).*

Number of experiments	n cycle value	M1 value (number of frequent items)	a1 value (support)	P value (indicator of economic benefits)
1	13	273	0.7988	1.402
2	21	69	0.9098	3.200

4.2. Experimental Analysis

The essence of this article is to hope to find suitable parameters through experiments. Given a certain data size, different economic benefit indicators have different support values for the P value, but in general, the higher the P value, the higher the support value., And the number of frequent episodes will decrease accordingly. Therefore, when making a decision, the actual economic benefit value should be fully considered. Too low will generate too many frequent sets, while too much will generate too few frequent sets and fewer association rules. The economic benefit values in Table 3 and Table 1 are set to be the same to better illustrate that the generation of association rules depends on the generation of frequent itemsets. Higher credibility only reflects the degree of dependence between commodities (sets), but this requires a certain data size to be more meaningful. Therefore, in order to further explain the relationship between the P value (economic benefit value) and various parameters, additional simulation experiments are set up. [9, 11, 12]

Additional simulation experiments

Experimental conditions: N=10, Nsum=100W, simulated experimental data 10 times.

Experimental purpose: To observe the relationship between P value and a1 parameter.

Experimental results: Figure 1 below.

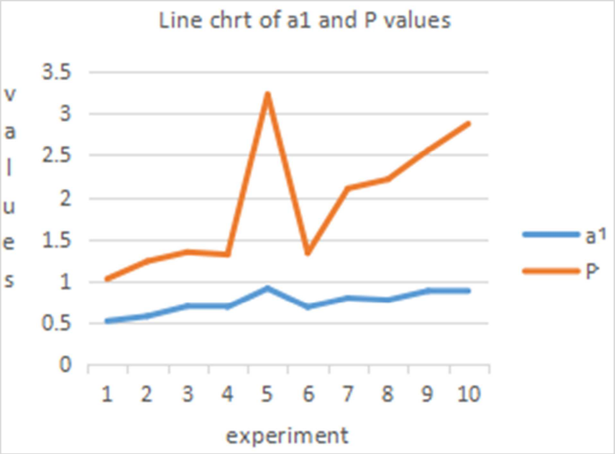

Figure 1. *Trend of a1 value and P value in 10 experiments.*

It can be seen from the figure that the P value trend is basically the same as the a1 value trend, but the M1 and M2 values in the tables are opposite. This requires decision-makers not to blindly pursue the P value and ignore the number of association rules. The value is in the acceptable range to pursue the P value to determine the values of a1 and a2.

5. Summary and Outlook

What this article shows is that we hope to find the appropriate support and credibility values in the association analysis. The choice of parameters in this article more considers the impact of the P (economic benefit) value, which is also in line with social psychology. But for decision makers, it is not completely dependent on it, there will be external secondary factors, so the conclusions in this article can be used as reference value [10].

In the binary search experiment, we have a hypothetical problem, that is, when the P value is less than or equal to 1 (benefit loss or equal), the parameter value (mainly the con or sup value) is considered to be too low and it is adjusted upward; On the contrary, the larger the P value, the larger the corresponding parameter value is and lowered appropriately. The establishment of this hypothesis is the premise of the experiments in this paper. In this article, there are some custom parameter values. For example, the P value is determined based on the relationship between the predicted economy and the decision cost. Different companies will have different methods of recognizing the P value [13]. P value is calculated by assigning inherent value weight to each frequent itemset. For example, the definition of L value is the degree of stability of the parameter values. In the experiment, L=0.01. The purpose of setting this parameter is to be able to produce stable parameter values, and it is also an exit condition set to facilitate the experiment without generating an infinite loop.. Therefore, to explain the problem of parameter selection strategy, we must fix certain parameters to better explain the conclusion [14, 15].

In order to achieve stable economic benefits, the conclusion of this paper is that it is hoped that experimental algorithms will be used to show that decision-making pursues the P value while not ignoring the number of association rules. The parameter value should be set within the acceptable range of the M value in this paper. There are still many problems to be solved in this paper, such as whether the hypothetical premise is really appropriate, the reasonable strategy of L value, and the various effects of data size on parameters will be problems that need to be solved. I hope this paper can bring to other studies Enlightenment and help.

References

[1] Wang Shuang, Yang Guangming, Zhu Zhiliang. Frequent item query algorithm based on uncertain data [J]. Journal of Northeastern University (Natural Science). 2011 (03).

[2] Yan Yuejin, Li Zhoujun, Chen Huowang. Efficient mining of maximum frequent itemsets based on FP-Tree [J]. Journal of Software. No. 2, 2005.

[3] Feng Yucai, Feng Jianlin. Incremental Updating Algorithm for Association Rules [J]. Journal of Software. 1998 (04).

[4] Song Yuqing, Zhu Yuquan, Sun Zhihui, Chen Geng. Algorithm for Mining and Updating the Maximum Frequent Itemsets Based on FP-Tree [J]. Journal of Software. 2003 (09).

[5] Cui Haili, Yuan Zhaoshan. A mining algorithm to quickly find the maximum frequent itemsets [J]. Journal of Hefei University of Technology (Natural Science Edition). 2006 (11).

[6] Wang Jinmiao, Zhang Longbo, Yan Guanghui, Wang Fengying. A Method for Mining the Maximum Frequent Itemsets in Uncertain Data [J]. Journal of Shandong University of Technology (Natural Science Edition). 2013 (05).

[7] Ma Lisheng, Deng Huiwen, Qi Yi. Mining algorithm of maximum frequent itemsets based on FP-tree [J]. Computer Engineering and Design. 2008 (02).

[8] Liu Junqiang, Sun Xiaoying, Wang Xun, Pan Yunhe. A New Method for Mining Maximum Frequent Patterns [J]. Chinese Journal of Computers. 2004 (10).

[9] Li Xiaoqing. Analysis and mining of customer association risk based on big data [J]. The era of financial technology. 2020 (04).

[10] Sun Keke; Li Zhong; Li Haiyang; Li Ying; Wang Yuanyuan University Library Access Control Data and Results Association Analysis [J]. Computer Knowledge and Technology. 2020 (02).

[11] Xiang Jianfeng, Research on the Architecture of Network Security Situation Awareness Platform Based on Big Data [J]. Science and Technology Innovation. 2020 (02).

[12] Wang Xiangrui. Application research of association rules mining in data mining technology [J]. Coal Technology. 2011 (08).

[13] Shen Yi, Wang Shuwang. Research on the mining of quantitative association rules [J]. Computer and Information Technology. 2005 (05).

[14] Li Xucheng, Wang Baobao. An improvement of Apriori algorithm in mining association rules [J]. Computer Engineering. 2002 (07).

[15] Zheng Lin. A divide-and-conquer Apriori algorithm that directly generates frequent itemsets [J]. Computer Applications and Software. 2014 (04). Biography.

Biography

Yan Yuanhai (1985-), male, native of Ji'an, Jiangxi, Guangdong University of Finance and Economics, School of Data Science, master degree, lecturer, mainly engaged in data visualization and data analysis algorithm research. Fund Project: 2017 "Innovative Strong School Project" Project Number: 2017KQNCX266.