A novel interestingness measure based on fusion model for association rules mining

Junrui Yang*, and *Lin* Xu

College of Computer Science and Technology, Xi'an University of Science and Technology, 710600 Xi'an, China

Abstract. Aiming at the shortcomings of the traditional "support-confidence" association rules mining framework and the problems of mining negative association rules, the concept of interestingness measure is introduced. Analyzed the advantages and disadvantages of some commonly used interestingness measures at present, and combined the cosine measure on the basis of the interestingness measure model based on the difference idea, and proposed a new interestingness measure model. The interestingness measure can effectively express the relationship between the antecedent and the subsequent part of the rule. According to this model, an association rules mining algorithm based on the interestingness measure fusion model is proposed to improve the accuracy of mining. Experiments show that the algorithm has better performance and can effectively help mining positive and negative association rules.

1 Introduction

Association rules mining is an important research direction in the field of data mining, and it has been widely used in financial, meteorological, medical and other fields[1]. With the development of technology, data in various fields have shown explosive growth. Using the traditional "support-confidence" mining framework to mine transactions in big data is not ideal. Both the practicality of the mined rules and the efficiency of the algorithm are obviously insufficient. For example, when using the traditional framework to mine association rules, usually only positive association rules can be mined, while the influence of negative association rules[2] on decision support is ignored.

According to the mining problems of positive and negative association rules[3] and the shortcomings of traditional mining frameworks, some researchers have proposed the concept of interestingness measure[4]. The interestingness measure indicates the degree of user attention to the rules, and this attention is a comprehensive consideration including the practicality, novelty and understandability of the rules. At present, interestingness measure is mainly divided into objective interestingness measure and subjective interestingness measure. Paper 5 solved the problem of asymmetry of positive and negative rules in mining by using correlation analysis and introducing negative confidence[5]. Literature 6 obtains the interestingness measure rules by calculating and comparing information entropy[6]. Literature 7 introduces the degree of related interest to measure the relevance between items in the

* Corresponding author: xltt1211@163.com

itemset[7]. According to the nature of the correlation measure, Paper 8 summarized an interestingness measure combined with confidence[8]. Literature 9 uses the difference probability interestingness measure to estimate and predict the weight of time series transactions to realize the streamlining of mining[9].

Since the objective interestingness measure does not consider the subjective influence of users and can better reflect the actual situation, most of the research on the interestingness measure model is carried out on the objective interestingness measure[10].Based on this, this paper proposes an objective interestingness measure, which is based on the difference idea interestingness measure[3], combined with the cosine measure, so that the interestingness measure can effectively express the relationship between the antecedent and the subsequent part of the rule, and the interestingness measure is used to solve the problem of mining positive and negative association rules.

2 Common interestingness measures

2.1 Lift measure

Lift[11] refers to the ratio between the confidence c of the rule and the subsequent parts of the rule. It represents the interrelationship between the antecedents and the subsequent parts of a rule.

$$lift(X \Rightarrow Y) = \frac{c(X \Rightarrow Y)}{s(Y)} = \frac{s(X \Rightarrow Y)}{s(X)s(Y)} = \frac{P(X \Rightarrow Y)}{P(X)P(Y)} \tag{1}$$

From formula(1), when lift(X⇒Y)>1, it means that the rule antecedent X and the rule subsequent part Y are positively correlated; when lift(X⇒Y)<1, it means the rule antecedent X and the rule subsequent part Y is negatively correlated; when lift(X⇒Y)=1, it means that the rule antecedent X and the rule subsequent part Y are independent of each other, and the rule antecedent has no effect on the rule subsequent, and vice versa.

2.2 IS measure

IS measure[12] is equivalent to the cosine value in mathematical knowledge. It is improved by adding mathematical concepts on the basis of lift, and mainly deals with asymmetric binary variables.

$$IS(X \Rightarrow Y) = \frac{s(X \Rightarrow Y)}{\sqrt{s(X)s(Y)}} = \frac{P(X \Rightarrow Y)}{\sqrt{P(X)P(Y)}} \tag{2}$$

The value range of the IS measure and the nature of judging positive and negative association rules are the same as the promotion. However, compared with the lift measure, the problem with the IS measure is that when the rule antecedent is negatively correlated with the rule's subsequent part, the metric value is very large. It is difficult to judge whether it is a true negative correlation.

2.3 PS measure

The PS measure[13] is a interestingness measure obtained from the definition of relevance through the probability method. It is defined as follows:

$$PS(X \Rightarrow Y) = s(X \Rightarrow Y) - s(X)s(Y) = P(X \Rightarrow Y) - P(X)P(Y) \tag{3}$$

The value range of the PS measure is [-0.25,+0.25]. When PS(X⇒Y)>0, it means that the pair of rules X⇒Y is a positive association rule, When PS(X⇒Y)=0, it means that X and Y are independent of each other, When PS(X⇒Y)<0, it means that X⇒Y is a negative association rule.

3 Objective-based interestingness measure fusion model

In the objective interestingness measure, through the analysis of some commonly used interestingness measures, and the objective-based interestingness measure introduced below, cleverly combined with the cosine measure, an interestingness measure fusion model is obtained, which effectively solves the original problem. There are defects in the model, and it can effectively mine positive and negative association rules.

The interestingness measure model based on the difference idea[14] is mainly different from other objective interestingness measures in that it adds the concept of confidence to improve the credibility of rule interest.

$$Interest(X \Rightarrow Y) = \frac{c(X \Rightarrow Y) - s(Y)}{max\{c(X \Rightarrow Y), s(Y)\}} \tag{4}$$

Among them, $\max\{c(X \Rightarrow Y), s(Y)\}$ is called the standardization factor.

The value range of this formula is [-1,+1]. The closer its value range is to 1, the stronger the positive relationship between X and Y, and the closer to -1, the stronger the negative relationship. The closer to 0, the weaker the correlation between X and Y.

The interestingness measure model based on the difference idea introduces confidence and standardization factors, but the instability of the result of the standardized factor in the denominator of the interest degree will cause the interest to be too large or too small, and it will also miss some rules or get some meaningless rules. By combining the cosine measure, the interestingness measure model based on the difference idea is improved, and the new interestingness measure model is obtained as follows:

$$Int(X \Rightarrow Y) = \frac{c(X \Rightarrow Y) - s(Y)}{\sqrt{s(X)s(Y)}} \tag{5}$$

The formula is equivalently transformed into:

$$Int(X \Rightarrow Y) = \frac{\frac{P(X \Rightarrow Y)}{P(X)} - P(Y)}{\sqrt{P(X)P(Y)}} = \frac{P(X \Rightarrow Y) - P(X)P(Y)}{P(X)\sqrt{P(X)P(Y)}} \tag{6}$$

Because $P(X)\sqrt{P(X)P(Y)} > 0$, and the numerator in the interestingness measure model is equal to the value of the PS measure, it proves that the interestingness measure can not only meet the feasibility of mining positive and negative association rules, And also has the characteristics of PS measure.

The interestingness measure model has the following properties:

Property 1 For rule X⇒Y, Int(X⇒Y) is not equal to Int(Y⇒X), that is, it distinguishes the difference between rule X⇒Y and rule Y⇒X, and distinguishes between transaction X and transaction Y characteristic.

For example. In a transaction set, where the total number of transactions is 1000, itemset X includes 500 transactions, itemset Y includes 300 transactions, and the total number of transactions including X and Y is 300. According to formula (5), Int(X⇒Y)≈0.76, Int(Y⇒X) ≈0.97.

Property 2 Int(X⇒Y)>0, X and Y are positively correlated, Int(X⇒Y)<0, X and Y are negatively correlated.

Proof: According to formula (6), the numerator of the interestingness measures model $P(X \Rightarrow Y) - P(X)P(Y)$ is equal to the PS measure, and $P(X)\sqrt{P(X)P(Y)} > 0$, so it satisfies the property of PS measure, so this property can be obtained.

Property 3 Set the support of the rule antecedent X and subsequent part Y to s=0.5, when s(X⇒Y)=0, then Int(X⇒Y)=-1, the negative correlation of the rule is the largest.

Property 4 Set the support of the rule antecedent X and subsequent part Y to s=0.5, and s(X⇒Y)=0.5, at this time Int (X⇒Y)=1, the positive correlation of the rule is the greatest.

Definition 1 The best support S. When the support of the rule antecedent X and the rule subsequent part Y is set to S, the value of |Int| reaches the maximum at this time, and S is called the best support at this time.

4 Association Rules Mining Algorithm Based on Interestingness measure

According to the interestingness measure model proposed in the previous section, combined with the support-confidence framework, an improved algorithm IntRIMine is proposed on the basis of the FP-growth algorithm, which improves the performance of the original algorithm and enables the algorithm to effectively mine strong association rules. And can dig out the positive and negative association rules respectively according to the interestingness measure model.

The IntRIMine algorithm has four main steps. First, the items in the dataset D are sequentially added to the FP-tree, and the support and confidence of the itemset are obtained according to the traversal of the tree. Then, the minimum support threshold is used to determine whether it is frequent itemsets, the obtained frequent itemsets are judged whether it is a strong association rule according to the minimum confidence threshold. Finally, according to the established interestingness measure model, it is judged whether the interestingness measure of the itemset in the obtained association rule meets the minimum interestingness measure threshold. If the interestingness measure of the itemset is greater than the minimum interestingness measure threshold and greater than 0, then the itemset is added to the set of positive association rules. Conversely, if the degree of interest is less than 0, judge whether the opposite value of the degree of interest is greater than the minimum interestingness measure, if it is satisfied, judge whether the counterexample meets the minimum support and minimum confidence, and if it meets the conditions, it is added to the negative association rules set.

5 Experimental comparison and analysis

The hardware environment used in this experiment is Inter Inter Core i5 CPU @ 2.60GHz, 8GB memory, the program is written in Python language, and the Pycharm 2019 development environment is used to compile and run.

The experiment uses the Groceries data set for testing. The Groceries data set records the real transaction records of a grocery store for a month. This experiment mainly tested and analyzed the performance of FP-growth algorithm, MAPPN algorithm and IntRIMine algorithm in terms of running time and the number of association rules generated.

5.1 Algorithm running time comparison

In order to verify the performance of the IntRIMine algorithm, the minimum confidence threshold is 0.7, the minimum interestingness measure threshold is 0.3, and the minimum

support threshold is changed between 0.05 and 0.25, the runtime changes of the three algorithms are observed.

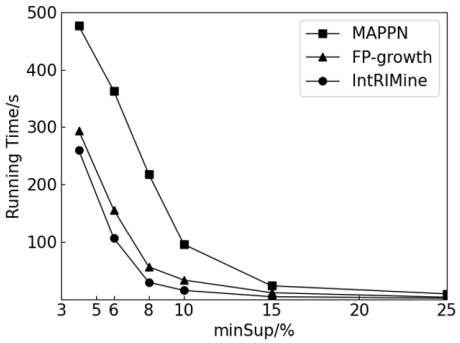

Fig. 1. The change of algorithm running time with support threshold.

It can be seen from Fig 1 that the running time of the IntRIMine algorithm is less than that of the FP-growth algorithm. This is because the IntRIMine algorithm is improved on the basis of the FP-growth algorithm, and the infrequent items in the candidate set are pruned from the tree in advance to improve Algorithm performance. In addition, because the MAPPN algorithm is based on the improvement of the Apriori algorithm, this type of algorithm consumes the most time because it needs to recursively generate frequent itemsets.

5.2 Comparison of the number of generated association rules

According to different confidence thresholds, different support thresholds and different interestingness measure thresholds, observe the comparison of the number of rules obtained by the algorithm IntRIMine using the Objective-based interestingness measure fusion model and the number of rules obtained by the traditional mining framework algorithm.

Table 1. Comparison of the number of generated association rules.

minConf	minSupp	minInt	association rules without interest	association rules using interest	
				positive correlation	negative correlation
0.6	0.1	0.3	2803	1759	831
		0.4		1526	745
	0.15	0.3	911	587	273
		0.4		462	254
0.7	0.1	0.3	1271	641	419
		0.4		533	395
	0.15	0.3	299	167	107
		0.4		142	95

interestingness measures model is less than using the traditional "support-confidence degree" framework, because the interestingness measure model fully considers the correlation between the rule antecedent and the rule subsequent. It can filter out some rules that are not really interesting to users using traditional frameworks, thereby improving the accuracy of mining. In addition, the number of rules mined is different according to the defined interestingness measures threshold. As the interestingness measure threshold increases, the number of rules mined will decrease. At the same time, compared with the traditional mining framework, using the interestingness measure model to mine association rules can not only mine the positive association rules, but also mine the hidden negative association rules.

6 Conclusion

This paper analyzes the commonly used interestingness measures, and merges the interestingness measure based on the difference idea and the cosine measure to obtain a new interestingness measure model. Use this model to mine association rules, so that it can mine positive and negative association rules effectively. By using a certain grocery store transaction dataset for verification, the algorithm in this paper is compared with the existing positive and negative association rules mining algorithm and the mining results using the traditional "support-confidence" framework. The results verify the effectiveness of the improved interest model and the algorithm using the model, and effectively improve the mining accuracy of the algorithm.

References

1. Y Cui, Z Q Bao. Survey of association rule mining[J]. Application Research of Computers, 2016, 33(02): 330-334.
2. N J Chen, Z N Gao. Improved Positive and Negative Association Rules Mining Algorithm[J]. Computer Science, 2011, 38(12): 191-193+212.
3. Y F Zhang, Z Y Xiong, Peng Yan, et al. Association Rule Mining Method Based on Interest Measure with Positive and Negative Items[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(03): 407-411.
4. Shaikh, Mateen R., McNicholas. Standardizing interestingness measures for association rules[J]. Statistical Analysis and Data Mining, 2018, 11(6): 282-295.
5. J L Lu, S W Chen. Mining association rules based on correlation measure[J]. Journal of Zhejiang University(Science Edition),2012,39(03):284-288.
6. Z Jin, R J Wang. Interestingness Rule Mining Algorithm Based on Information Entropy[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(06):524-532.
7. Y Q Zhang, C Wang. Association rule mining algorithm based on related interest measure[J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science Edition),2017,37(05):87-93.
8. Y Q Ma, T Wu, X K Deng. Research on Mining Method of Positive and Negative Association Rules Based on Interestingness Measurement[J]. Computer Technology and Development, 2018, 28(05): 38-41+46.
9. Z F Wang, H J Zhao, Li Cong, et al. Mining Association Rules for Multi-Class Difference Data of Web Services Based on Interest Measure Function[J]. Computer Application and Software,2019,36(12):60-65+105.
10. M H Wang, S M Wu, R C Cai. Two novel interestingness measures for gene association rule mining[J]. Neural Computing and Applications, 2013, v 23, p 835-841.
11. Abdellatif, Safa , M. A. B. Hassine , and S. B. Yahia. Novel Interestingness Measures for Mining Significant Association Rules from Imbalanced Data[J]. The Workshops of the 33rd International Conference on Advanced Information Networking and Applications WAINA-2019, 2019, v 927, p 172-182.
12. B H Liang, M Cai. Improvement of Positive and Negative Association Rule Mining Method and Its Application[J]. Computer Engineering, 2010, 36(16): 44-46.
13. H F Zhou, Y Y Zhu. A Mining Algorithm for Association Rules based on Interest Measure[J]. Journal of Computer Research and Development, 2002(04): 450-457.
14. W W Wang, X F Xia, X M Li. Personal interest degree model based on consumer behavior[J]. Computer Engineering and Applications, 2012, 48(08): 148-151+199.