Single Versus Multiple Measures for Fuzzy Association Rule Mining

Keerachart Suksut*, Phaichayon Kongchai, Sak Phoemhansa,
Rattaphong Sutamma, Kittisak Kerdprasop, Nittaya Kerdprasop

Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000 THAILAND.

*Corresponding Author: mikaiterng@gmail.com

Abstract

Data mining can identify patterns of data, find relationships within the data to predict the outcome or predict future data trends. Association rule mining is part of data mining and it has been applied in many fields. The performance of association rule mining depends on the support and confidence measures. In this research, we perform a comparative of number of rules that is affected from different kinds of measures. We propose the idea of using a combination of measures (such as support*confidence, confidence*lift) instead of considering only support value or confidence. The main objective is to reduce the number of rules in the process of fuzzy association rule mining, in which a compact rule set is important to the real application of the fuzzy system. The experimental results reveal that our idea of combining multiple measures for fuzzy association rule mining can significantly reduce the number of association rules.

Keywords: Data mining, Association Rule, Fuzzy Association rule mining, Measure for Association Rules.

1. Introduction

At present, data mining has been very popular because it helps extract, search, predict, and many other knowledge-intensive tasks. Usefulness is from employing existing data to predict future trends in data, extracting patterns of the data, and finding the relationship within the data group. These benefits make data mining been widely applied. In medicine, it is used in the data analysis of the patient to see a trend whether or not the patient is likely to be ill from a disease. The business side also uses data mining to analyze the buying patterns of consumers.

Data mining techniques can be practiced in various ways. The technology that has been adopted widely is association rule mining in relation to the incident and bring those things that occur together to create the association rules.

Finding association rules was proposed by Agrawal⁽¹⁾ to be used in data mining. To find efficiently association rules a criterion for calculating the support of rule is necessary. A confidence criterion is also needed to reduce the number of rules.

The most relevant research to ours is to reduce the number of rules in different ways, such as the work of F. P. Pach et al.⁽²⁾. They proposed the concept to reduce the number of fuzzy association rules with pruning technique, which makes the association rules smaller in their size. The research of M. M. Ballesteros et al.⁽³⁾ aimed to improve the algorithm and they compared the fitness function obtained from the various developed algorithms as well as there is quite a few the traditional algorithm. However, research work that compares the number of rules based on different measures, especially a multi-criteria measure. Therefore, we propose this research that was performed to compare the number of rules for different measures and also apply more than one measures to reduce the number of discovered association rules.

2. Background

2.1 Association Rule Mining

Association rule mining is a process that gained much popularity in data mining. The process to find relationships that hidden in a dataset can be explained with the following example.

From Table 1 which is a set of ten transactions, the

frequency of the purchases for each product type in association to other products are shown in table 2.

From Table 2, one can create an instance of the relationships that can be represented as “if antecedent then consequence” (or “antecedent $\Rightarrow$ consequence”) as follows:

If customer buy milk, then they also buy bread.

If customer buy beer, then they also buy potatoes.

If customer buy bean, then they also buy potatoes.

Table 1 Transactional database of customer’s purchases.

Transaction ID	Item
1	Milk, Bread, Butter, Water
2	Milk, Bread, Cola
3	Beer, Bean, Potatoes
4	Milk, Bread, Water
5	Water, Cola, Beer
6	Milk, Butter, Potatoes
7	Cola, Potatoes, Bean
8	Milk, Bread
9	Milk, Bread, Potatoes
10	Beer, Bean, Potatoes

Table 2 Frequency of each combination of two product types.

	Mil k	Brea d	Butte r	Wate r	Col a	Bee r	Bea n	Potatoe s
Milk	6*	5	2	2	0	0	0	2
Bread	5	5*	1	2	1	0	0	0
Butter	2	1	1*	1	0	0	0	1
Water	2	2	1	3*	1	1	0	0
Cola	0	1	0	1	3*	1	1	1
Beer	0	0	0	1	1	3*	2	2
Bean	0	0	0	0	1	2	3*	3
Potatoes	2	0	1	0	1	2	3	4*

*Purchases counted as one-item frequency.

Table 3 Measures for finding support of rules.

Measure	Equation	Range
support(X)	$n(X) / N$	[0,1]
support($X \Rightarrow Y$)	$n(X \cap Y) / N$	[0,1]
confidence($X \Rightarrow Y$)	$support(X \Rightarrow Y) / support(X)$	[0,1]
lift($X \Rightarrow Y$) ⁽⁴⁾	$support(X \Rightarrow Y) / (support(X) * support(Y))$	[0,∞]
conviction($X \Rightarrow Y$) ⁽⁵⁾	$(1 - support(Y)) / (1 - confidence(X \Rightarrow Y))$	[0,∞]
gain($X \Rightarrow Y$) ⁽⁶⁾	$confidence(X \Rightarrow Y) - support(Y)$	[-0.5,1]
leverage($X \Rightarrow Y$) ⁽⁷⁾	$support(X \Rightarrow Y) - (support(X) * support(Y))$	[-0.25,0.25]

2.2 Different Measures of Association Rule Mining

(a) Support measure

Support measure is frequency of each purchased item. This measure supports usefulness of association rules. The higher the frequent that association occur. The boundary of support measure ranges between 0 to 1 and can be computed as in the following equation:

$$support(X \Rightarrow Y) = \frac{n(X \cap Y)}{N} \quad (1)$$

where:

$n(X \cap Y)$ is the number of time that the item X happened together with Y in the same transaction.

N is the total number of transaction.

Example 1: finding the value of support of rule with the support measure.

From Table 1 when using the support measure to find the value of support of rule that customers buy milk and also buy bread, the computation is as follows:

$$support(milk \Rightarrow bread) = \frac{5}{10} = 0.5$$

That means the value of support of rule “if milk then bread” is 0.5, or 50% of the whole transactions.

(b) Confidence measure

Confidence measure is the measure to indicate the reliability of the rule. The boundary of confidence measure is between 0 to 1 and can be computed as follows:

$$confidence(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X)} \quad (2)$$

where:

support($X \Rightarrow Y$) is the frequency that both items X and Y occur in the same transaction,

support(X) is the support that X has happened counted from all transactions.

Example 2: finding the value of confidence of rule with the confidence measure.

From Table 1 when using the confidence measure to find the value of confidence of rule that customers buy milk and also buy bread, the computation is as follows:

$$confidence(milk \Rightarrow bread) = \frac{0.5}{0.6} = 0.833$$

That means the value of confidence of rule that if milk is bought then bread also be bought is 0.833, or 83% of confidence.

(c) Lift measure

Lift measure has its boundary of value between 0 to infinity. If the value is less than 1, it means that X and Y are related in a negative way. If a value equal 1, it means that X and Y are independent. If a value is greater than 1, it indicates that X and Y are correlated in a positive way. The computation is shown in the following equation:

$$lift(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X) * support(Y)} \quad (3)$$

where:

support($X \Rightarrow Y$) is the frequency that X and Y occur in the same transaction,
support(X) is the support that X happened against all transactions,
support(Y) is the support that Y happened against all transactions.

Example 3: finding the value of lift of rule with the lift measure.

From Table 1 when using the lift measure to find the lifted confidence that customers who buy milk also buy bread, the lift value computation can be illustrated as follows:

$$lift(milk \Rightarrow bread) = \frac{0.5}{0.6 * 0.5} = 1.667$$

That means the value of lift of rule $milk \Rightarrow bread$ is 1.667; that is, milk and bread have a positive relationship.

(d) Conviction measure

Conviction measure has its boundary value between 0 to infinity. If the value is less than 1, it means that X and Y are related in a negative way. If a value equal 1, it means that X and Y are independent. If a value is greater than 1, it indicates that X and Y are correlated in a positive way. The conviction measure can be computed as in the following equation:

$$conviction(X \Rightarrow Y) = \frac{(1 - support(Y))}{(1 - confidence(X \Rightarrow Y))} \quad (4)$$

where:

support(Y) is the support value that Y happened against all transactions,
confidence($X \Rightarrow Y$) is the reliability of the X and Y co-occurring in the same transaction.

Example 4: finding the value of conviction of rule with the conviction measure.

From Table 1 when using the conviction measure to find the conviction of rules that customers who buy milk also buy bread, the computation is as follows:

$$conviction(milk \Rightarrow bread) = \frac{(1 - 0.5)}{(1 - 0.833)} = 2.99$$

That means the conviction of rule that if milk is bought then bread is also bought is 2.99. That means milk and bread have a positive relationship.

(e) Gain measure

Gain measure is a measure that is used to compute added value or change of rules. The boundary of gain measure is between -0.5 to 1. Its computation is as follows:

$$gain(X \Rightarrow Y) = confidence(X \Rightarrow Y) - support(Y) \quad (5)$$

where:

support(Y) is the support that Y happened against all transactions,
confidence($X \Rightarrow Y$) is the reliability of the X and Y co-occurring in the same transaction.

Example 5: finding the gain value of rule with the gain measure.

From Table 1 when using the gain measure to compute usefulness of rules that customers who buy milk also buy bread, the computation is as follows:

$$gain(milk \Rightarrow bread) = 0.833 - 0.5 = 0.333$$

That means the gain value of rule that if milk is bought then bread is also bought is 0.333.

(f) Leverage measure

Leverage measure is a measure that indicates the strength of the rule. If the leverage is less than 0, it means that X and Y are related in a negative way. If the value is equal to 0, it means that X and Y are independent, and any value greater than 0 indicates that X and Y are positively dependent. The computation is as follows:

$$leverage(X \Rightarrow Y) = support(X \Rightarrow Y) - (support(X) * support(Y)) \quad (6)$$

where:

support($X \Rightarrow Y$) is the support that X and Y co-occurring in the same transaction,
support(X) is the support that X happened against all transactions,
support(Y) is the support that Y happened against all transactions.

Example 6: finding the value of leverage of rule with the leverage measure.

From Table 1 when using the leverage measure to evaluate the rule $\text{milk} \Rightarrow \text{bread}$, the computation is as follows:

$$\text{leverage}(\text{milk} \Rightarrow \text{bread}) = 0.5 - (0.6 * 0.5) = 0.2$$

That means the value of leverage of rule that if milk is bought then bread is also bought is 0.2, which implies milk and bread have a positive relationship.

Figure 1 A research framework for comparing the effect of different measure for fuzzy

Figure 2 Frequent fuzzy item set algorithm⁽²⁾

Figure 3 Fuzzy association rule-base generation algorithm⁽²⁾

3. Methodology

Researchers have designed the process and findings of the comparison of single measure and multi-measure for the discovery of fuzzy association rules as show in Figure 1.

From figure 3 can explain the process of the research are as follows:

- Input data partitions

In the process of the associative-based classification is the partitioning of an attribute whose value is continuous. It is a transformation of continuous values to be discrete intervals based on the concept of fuzzy sets.

- Frequent fuzzy item set generation

At this process, we will search and count the frequency of items and combination of items. These items are to be used in the creation of rules, regardless of the support of the rule. The procedure is shown in the Figure 2.

- Fuzzy association rule generation

In this process, we will create the rules that are screened by different measures and will vary various minimum support thresholds in order to create various final rule sets. The procedure works as shown in figure 3.

4. Experimental Results

This study used Breast Cancer Wisconsin data that are obtained from the UCI Machine Learning Repository. This dataset has 699 instances with 10 attributes and a target attribute of two classes. We used this data to find fuzzy association rules with various single measures and different combinations of two measures. The support of the rules has been set to be 90%, 80%, 70% 60%, and 50%. The setting of each measuring value is summarized and shown in Table 4. The experimental results are shown in Tables 5-7.

The results shown in Table 5 are the comparison of number of rules obtained from using different single measure with the support of rules ranging from 90%, 80%, 70%, 60%, and 50%. It can be seen that the each single measure gives a number of fuzzy association rules different from the value in each criteria. There are some measures of screening fuzzy association rules that appeared to produce rules less than the minimum support of the rule. We noticed that the number of rules in each measure is not equal.

Table 4 Percentage value for each measure used in the experimentation compared with the boundary value for each measure.

	50%	60%	70%	80%	90%	100%
support	0.5	0.6	0.7	0.8	0.9	1.0
confidence	0.5	0.6	0.7	0.8	0.9	1.0
lift	0.5	0.6	0.7	0.8	0.9	1.0
conviction	0.5	0.6	0.7	0.8	0.9	1.0
gain	0.475	0.58	0.685	0.79	0.895	1.0
leverage	0	0.05	0.1	0.15	0.2	0.25

Table 5 Number of rules and average values obtained of each single measure (SOR is minimum support of rule).

		Measure					
		Support	Confidence	Gain	Leverage	Lift	Conviction
SOR = 90%	Number of Rules	0	105	0	0	252	21
	Average	0	0.9392	0	0	1.327	2.0459
SOR = 80%	Number of Rules	0	105	0	84	252	21
	Average	0	0.9392	0	0.1657	1.327	2.0459
SOR = 70%	Number of Rules	0	238	0	210	252	21
	Average	0	0.8801	0	0.1487	1.327	2.0459
SOR = 60%	Number of Rules	22	252	0	252	252	21
	Average	0.6293	0.8680	0	0.1351	1.327	2.0459
SOR = 50%	Number of Rules	252	252	0	252	252	21
	Average	0.5651	0.8680	0	0.1351	1.327	2.0459

Table 6 Number of rules and average values obtained from a combination of support and other five criteria (SOR is minimum support of rule).

		Measure				
		Support* Confidence	Support* Gain	Support* Leverage	Support* Lift	Support* Conviction
SOR = 70%	Number of Rules	0	0	8	216	21
	Average	0	0	0.1009	0.7630	1.2289
SOR = 60%	Number of Rules	4	0	210	252	21
	Average	0.625	0	0.0831	0.7479	1.2289
SOR = 50%	Number of Rules	108	0	252	252	21
	Average	0.5356	0	0.0759	0.7479	1.2289

Table 7 Number of rules and average values obtained from a combination of confidence and other criteria (SOR is minimum support of rule).

		Measure				
		Confidence * Support	Confidence * Gain	Confidence * Leverage	Confidence * Lift	Confidence * Conviction
SOR = 70%	Number of Rules	0	0	208	252	21
	Average	0	0	0.1307	1.1537	1.4265
SOR = 60%	Number of Rules	4	0	234	252	21
	Average	0.6250	0	0.1236	1.1537	1.4265
SOR = 50%	Number of Rules	108	0	208	252	21
	Average	0.5356	0	0.1180	1.1537	1.4265

When using more than one measures for determining the number of fuzzy association rules, we set the support of the rule from 70% were not able to give any rule.

It can be seen from Table 6 that a comparison in terms of of number of rules obtained from using different 2 measures by using support measure as the main measure and varying support of the rule from 70% down to 60% and 50% can produce a small set of fuzzy association rules. Using more than a single measure as the condition in the selection of rules will result in less fuzzy association rules because the support of rules when combined with other calculation criteria results in the decrease of the rules that satisfy the threshold.

When we use more than a single measure to screen rules using the confidence as the main measure with the support of the rule ranging from 70%, 60%, to 50%, the results are shown as in Table 7.

It can be seen that the number of fuzzy association rules by using confidence measure as the main measure is different from using support measure as the main measure. Using the confidence as the main measure can provide a slightly larger number of fuzzy association rules than the use of support measure as the main measure. This may be can see from the

fact that confidence is a more relax criteria than the support measure.

5. Conclusions

Association rule mining is one technique of data mining that has been widely used in a number of applications. For some numeric and continuous values, mining for association rules cannot be applied directly because of the enormous amount of possible values. Fuzzy set concept can be applied to handle this situation and gives rise to the new sub-area called fuzzy association rule mining. A small number of rules given by the fuzzy association rule mining process is important to the performance justification of the process. In the work we propose the findings from our experimentation that a combination of measures to select the final rule set gives a better result than a single measure. Moreover, we found that the two-measures such as support*lift and confidence*leverage yield a reasonable set of fuzzy association rules.

References

- (1) Rakesh Agrawal, Tomasz Imielinski and Arun Swami. Mining Association Rules between Sets of Items in Large Database. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993.
- (2) Ferenc Peter Pach, Attila Gyenesei and Janos Abonyi. Compact Fuzzy Association Rule-Based Classifier. Expert System with Applications. Vol. 34, Issue 4, pp. 2406-2416, 2008.
- (3) Maria Martinez Ballesteros, Francisco Martinez Alvarez, Alicia Troncoso and Jose C. Riquelme. Selecting the Best Measures to Discover Quantitative Association Rules. Neurocomputing. Vol. 126. pp. 3-14, 2014.
- (4) Sergey Brin, Rajeev Motwani Motwani and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. Proceedings of the ACM SIGMOD. Vol. 26. pp. 265–276, 1997.
- (5) Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. Proceedings of the ACM SIGMOD. pp. 265–276, 1997.
- (6) Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: a survey. ACM Computing Surveys. Vol. 38. Issue 3. pp. 1–42, 2006.
- (7) G Piatetsky Shapiro. Discovery. Analysis and presentation of strong rules. Knowledge Discovery in Databases. pp. 229–248, 1991.