A boolean based method for evaluating relationships of noise tolerant itemsets in binary transactions

Morrel V.L.Nunsanga **Somen Debnath** **R.Lalchhanhima**

***Abstract*— Association rule mining deals with extracting important information hidden in large volumes of data and the standard metrics in the rule shows the relationship among the itemsets. There are different approaches which can be used to find interesting patterns or frequent patterns from a large pool of data. Support and Confidence are the standard metrics to measure the interestingness of the generated patterns. The tradition methods of evaluating these two standard metrics sometimes do not work well when the collected pool of data contains errors or noise in it. This paper proposed an approach for evaluating values of standard metrics of the frequent patterns generated from the pool of data which is applicable to the noise tolerant itemsets.**

***Index Terms*— Association rule, Frequent Itemsets, Noise Tolerant frequent itemsests, confidence, support.**

I. INTRODUCTION

Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. **Do not change the font sizes or line spacing to squeeze more text into a limited number of pages.** Use italics for emphasis; do not underline.

Uses of computer technologies in our daily activities and business patterns, such as e-commerce, web logs, financial transactions, data in research, computerization in organizations, etc generate huge amount of data in a short period of time, and these activities increase rapid growth in the storage of information and data in large volume. These large amount of data become a good source of useful information for researchers which in turn will help in making a good decision in business policy. Frequent pattern mining plays an essential role in mining associations[1,4]

A key task of association rule analysis is finding frequent itemsets, which are sets of items that frequently occur together in a transaction [3]. An example of such a rule might be that 90% of customers who purchase bread would purchase butter also. Finding such kind of relationship among items is valuable for making business plan, catalogue design, store layout, customer segmentation based on buying patterns. The database involved in this application is very large.

Many research studies on mining interesting patterns, such as [1, 3, 5, 7, 8,10], adopt an Apriori-like approach, which is based on an anti-monotone Apriori heuristic [1]: if any length k pattern is not frequent in the database, its length (k +1) super-pattern can never be frequent. This property can be applied to traditional mining approach where errors were not considered in the relationship of the itemsets.

In fact, in such a large pool of data, numerous errors or noise might appear in transactions due to different reasons such as items not properly recorded, or items were out of stock, or due to machine failures, etc. In the presence of such "fault" or "error" or "noise" the traditional frequent itemset algorithms may miss many useful patterns in the data set as it does not handles noise tolerant frequent itemsets. So, it is required to handles these kind of noises or errors in our dataset i.e it is required and highly necessary to consider noise itemsets in our dataset.

In the standard definition of frequent itemset, every transaction supporting the itemset must contain every item. But in case of noise tolerant frequent itemsets, it is enough that each transaction contains most of the items in the specified itemset. If we apply traditional approach to handles the noise tolerant itemsets in the large dataset, it may produce many unnecessary itemsets which may not at all interesting and it is difficult to have the actual measure of the relationship of the items in the itemsets.

C.Yang,U fayd et.al[2] introduced very interesting algorithms for noise tolerant frequent pattern mining. Studies on the properties of ETI also have been published in the paper which became the bench mark for the many algorithms. Many research works [6,9,15,17] have been done on mining interesting patterns in presence of noises or errors. In every association rule mining process, there are always two metrics such as support and confidence, by which the relationship or the interestingness of the itemsets are measured.

To summarize our contribution, we find that though the traditional approach for measuring confidence and support of frequent itemsets is applicable to noise tolerant itemsets in normal cases, this approach sometimes does not give

Manuscript received March, 2018.

***Morrel V.L.Nunsanga**, Department of IT, Mizoram University (e-mail: morel.hmar@mzu.edu.com). Aizawl, India, 9862596269*

***Somen Debnath**, Department of IT, Mizoram University (e-mail: somen.nit@gmail.com). Aizawl, India, 9774143253*

***R.Lalchhanhima**, Department of IT, Mizoram University (e-mail: chhana.mizo@gmail.com). Aizawl, India, 9612160719.*

meaningful confidence values of frequent itemsets. Some cases were highlighted so as to show these problems. A new framework has been proposed for measuring values of confidence and support of noise tolerant frequent itemsets which can also be applicable to traditional frequent itemsets. Solution to the stated problems are given with the proposed approach. The analysis of the proposed framework shows that our new approach always gives meaningful values of confidence and support of frequent itemsets in all cases.

II. BACKGROUND

Using notations, we briefly summarize the standard definition of association rule, support, confidence, noise tolerant itemset. Applying traditional confidence and support to noise tolerant itemsets.

Definition 1: Association rule[1]:

Let $I = (i_1, i_2, ...i_m)$ be a set of literals, called items. Let *D* be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form *X =>Y*, where $X, Y \subseteq I$, and $X \cap Y = \phi$; The sets of itemsets X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively [18]

Definition 2: Support

For a binary transaction data matrix D with transactions T and items I, the support of an itemset, i.e., a set of binary attributes, X ⊆ I, is given by[6]

$$\sigma(X) = \left|\{t \in T : D(t,i) = 1, \forall i \in X\}\right|$$

which is simply the count of transactions containing all the items of *X*.

The support for a particular association rule X => Y is the proportion of transactions in D that contain both X and Y.

$$Support\ (A \Rightarrow Y) = P(X \cup Y)$$

$$= \frac{\text{Number of transactions containing both X and Y}}{\text{Total number of transactions}}$$

Definition 3: Confidence

The confidence of the association rule $X \rightarrow Y$, where X ⊆ I, Y ⊆ I, and X ∩ Y = ∅ , is given by[6]

conf(X → Y) = σ (X ∪ Y)/ σ (X) = σ (X → Y)/ σ (X).

The confidence of the association rule X=>Y is a measure of the accuracy of the rule, as determined by the percentage of transactions in D containing X that also contain Y.

$$= \frac{\text{Number of transactions containing both X and Y}}{\text{Number of transactions containing X}}$$

Defintion 4: Noise Tolerant Itemset(NTI)

An itemset $E \subseteq I$ is a noise tolerant itemset having error ε and support k with respect to a database *D* having n transactions if there exists at least *k*n* transactions in which at least a fraction 1-ε of the items from E are present[2]

The fundamental difference between noise tolerant itemsets and traditional frequent itemsets is a relaxation in support criteria. In noise tolerant itemsets, the criterion for exact matching in normal itemsets is relaxed upto some threshold value to produce more flexible definition of support[2].

Table 1(a): sample dataset1

Transaction	Itemsets
T1	B,C,D,E,F
T2	A, D
T3	A, B, C, D, E
T4	A, C, F
T5	A, B, C, D, E, F
T6	A, C, D, F
T7	B, C, D, E, F
T8	A, B, C, D, E, F

Traditional Approach to Confidence for Noise Tolerant Itemsets.

Consider the dataset in table 1(a) sample dataset1, let us calculate the confidence for (X=>Y) for the dataset shown in table 1. Let X = {A,B, C, D}, Y={E, F}, and max_error ε = 0.25, k=50%, then the given sets are noise tolerant itemsets. From the above table, *Sup*($X \bigcup Y$) = 5, *Sup*(*X*) = 6

Conf(*X => Y*) =*Sup*($X \bigcup Y$) / *Sup*(*X*) = 5/6 = 0.83 or 83 %

Sup(X=>y) = *Sup*($X \bigcup Y$) / T = 5 / 8 = 0.625 or 62.5 %

III. PROBLEMS WITH THE EXISTING SYSTEM

Traditional approach for finding the value of support and confidence at times does not give a meaningful value.

Case 1: Using the dataset in table 2(a), *Let X = {A, B, C, D}, Y={E, F, G, H}*, k=50%(minimum support count), max_error ε = 0.25, then the given sets are fault tolerant itemsets.

From the table 2(a),

Sup($X \cup Y$) = 6, *Sup*(*X*) = 5

Conf($X \Rightarrow Y$) =*Sup*($X \cup Y$) / *Sup*(*X*)

= 6/5 = 1.2 or 120%

Sup(X⇒ y) = *Sup*($X \cup Y$) / T

= 6 / 9 = 0.66 or 66 %

The above result obtained is odd because the confidence value is more than 1 (ie more than 100%)

Table 2(a): sample dataset2

Transaction	Itemsets
T1	B, C, D, E, F, H
T2	A, D, G, H
T3	C, D, E, G
T4	A, C, E, F, G, H
T5	A, B, C, D, E, F,
T6	A, C, D, F, G, H
T7	B, C, D, E, F, G, H
T8	A, B, C, D, E, F, G, H
T9	B, D, F, H

Case 2: Consideing the dataset in table 3(a), suppose ε =0.5, k=50% (i.e. at least a transaction must contain 4/8 items and half of the transaction support the patter. If *X = {A, B, C, D}* and *Y = {E, F, G, H}*, then both *X* and *Y* are Noise tolerant Itemsets.

Table 3(a): Sample dataset

Transaction	Itemsets
T1	A, B, C, D
T2	A, B, C, D
T3	A, B, C, D
T4	A, B, C, D
T5	A, B, C, D
T6	E, F, G, H
T7	E, F, G, H
T8	E, F, G, H
T9	E, F, G, H
T10	E, F, G, H

Using traditional approach,

$Conf(X => Y) = Sup(X \cup Y) / Sup(X)$
= 10/5 =2 = 200%

As per the standard definition of the confidence, the result is very odd because:

i) Observing at the itemsets of both sides of the rule, the noise tolerant itemsets of antecedent never co-occurs with the noise tolerant itemsets in consequent, so, the confidence value is expected to be zero (0)
ii) The confidence value should always lie between 0 to 1

Thus, the traditional framework for measuring confidence of patterns does not seem appropriate for noise tolerant itemsets in many cases.

IV. PROPOSED APPROACH FOR CALCULATING SUPPORT AND CONFIDENCE

Consider the rule *(X=>Y)*, let '*T*' be the total number of transactions, max_error tolerant threshold ' ε ' and 'k' support threshold. In order to calculate the support and confidence of frequent itemsets, the following steps are proposed:

Step 1: Construction of Left binary array X_{xt} for Antecedent:

- If the Antecedent of the rule contains '0' more than the threshold value ε in a transaction, then assign 0 against that transaction in the X_{xt}, else assign 1. Repeat the same for all transactions.

Step 2: Construction of Right binary array Y_{yt} for Consequent

- If the Consequent of the rule contains noises more than the threshold value ε in a transaction, then assign 0 against that transaction in the Y_{yt}, else assign 1. Repeat the same for all transactions.

Step 3: ANDing X_{xt} and Y_{yt} support value of ($X \bigcup Y$) :

- AND the value of X_{xt} and Y_{yt}; increment the support count of the rule S_{xy} whenever the *ANDing* result is 1. Increment the support count S_x whenever the array X_{xt} contains 1.
-

Step 4: Calculate Confidence of the rule (X=>Y):
The final S_{xy} value is divided by the support count of the left side itemset S_x

$Conf(X=>Y) = S_{xy}/S_x$

Table 1(b) : dataset1

	I1	I2	I3	I4	I5	I6
T1	0	1	1	1	1	1
T2	1	0	0	1	0	0
T3	1	1	1	1	1	0
T4	1	0	1	0	0	1
T5	1	1	1	1	1	1
T6	1	0	1	1	0	1
T7	0	1	1	1	1	1
T8	1	1	1	1	1	1

Step 5: Calculate Support of the rule (X=>Y):
The final S_{xy} value is divided by the total number of transactions '*T*'

$Sup(X->Y) = S_{xy}/T$

V. DISCUSSION AND RESULT

We have introduced an approached to calculate the support count and the confidence value of a rule. The proposed approach is applicable to any binary transactions itemsets. The experimental result show that a meaningful results always been observed with the proposed approach

Consider the dataset1 from table 1 and let us rewritten the transactions in binary forms

X_{xt} = [1,0,1,0,1,1,1,1] , this implies S_x=6,
Y_{yt} = [1,0,0,1,1,1,1,1]
$S_{xy} = \sum(X_{xt}$ AND $Y_{yt})$
$= \sum$ ([1,0,0,0,1,1,1,1]),
Therefore, Support count S_{xy} = 5
We have $Conf(X=>Y) = S_{xy}/S_x$
= 5/6 = 0.83=83%
$Sup(X=>Y) = S_{xy}/T$ = 5/8 = 0.625 = 62.5%
We observed that in normal case, the proposed approach is giving the same value with the traditional approach.

Consider dataset2 shown in the table 2(a) and putting in binary form, we get the result as in table 2(b).
Let X = {A, B, C, D}, Y={E, F, G, H}, k=50%, max_error ε = 0.25, (as given in case 1)

X_{xt}=[1,0,0,0,1,1,1,1,0], this implies S_x = 5
Y_{yt}=[1,0,0,1,0,1,1,1,0]
Sup (X=>Y) = $\sum(X_{xt}$ AND $Y_{yt})$
$= \sum$ ([1,0,0,0,0,1,1,1,0])
Therefore Support count, S_{xy}= 4

We have,
Conf(X=>Y) = S_{xy}/S_x = 4/5 = 0.8=80%
Sup(X=>y) = S_{xy} / T = 4/9 =0.44 = 44%

The Confidence value of the itemsets 0.80(80%) calculated with the new approach here is comparatively meaningful than the previous result computed with traditional approach.

Table 2(b) : Sample dataset2

	A	B	C	D	E	F	G	H
T1	0	1	1	1	1	1	0	1
T2	1	0	0	1	0	0	1	1
T3	0	0	1	1	1	0	1	0
T4	1	0	1	0	1	1	1	1
T5	1	1	1	1	1	1	0	0
T6	1	0	1	1	0	1	1	1
T7	0	1	1	1	1	1	1	1
T8	1	1	1	1	1	1	1	1
T9	0	1	0	1	0	1	0	1

Table 3(b) : sample dataset3

	A	B	C	D	E	F	G	H
t1	1	1	1	1	0	0	0	0
t2	1	1	1	1	0	0	0	0
t3	1	1	1	1	0	0	0	0
t4	1	1	1	1	0	0	0	0
t5	1	1	1	1	0	0	0	0
t6	0	0	0	0	1	1	1	1
t7	0	0	0	0	1	1	1	1
t8	0	0	0	0	1	1	1	1
t9	0	0	0	0	1	1	1	1
t10	0	0	0	0	1	1	1	1

Solution to Case 2: Consider the dataset 3 shown in the table 3(a) and putting it in binary form table as in table 3(b).

If X = {A, B, C, D} and Y = {E, F, G, H}, where $\varepsilon = 0.5$, k=0.5,
By applying the same procedure, we have
S_x=4,
Sxy = $\sum(X_{xt}$ AND $Y_{yt})$
= $\sum$ ([0,0,0,0,0,0,0,0])
This implies S_{xy} = 0

Conf(X=>Y) = S_{xy}/S_x
= 0/4 = 0

The confidence value 0 obtained here is meaningful because, the left itemsets and the right itemsets never occur together.

The results obtained from different sampled data given are in table 4. From the result, we can see that values obtained with our new approach always provides intuitive and meaningful values for confidence and support in different situations.

Table 4 : Result analysis

		Tradition	Proposed Approach
Dataset1	Support	62.5%	62.5%
	Confidence	83%	83%
Dataset2	Support	66%	44%
	Confidence	120%	80%
Dataset3	Support	100%	0%
	Confidence	200%	0%

VI. CONCLUSION

The proposed approach here is a framework which is used to calculate confidence and support of an association rule in presence of noise. As experimented with different datasets, using the traditional approach, Confidence value of a frequent itemsets sometimes does not give the actual relationship among the items in the noise tolerant itemsets. It has been demonstrated that how this framework solves such problems and provide meaningful and intuitive value of confidence and support of the noise tolerant itemsets . The framework is still applicable to calculate the values of confidence and support of traditional frequent itemsets.

REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases*, VLDB. Vol. 1215. 1994.
[2] Yang, Cheng, Usama Fayyad, and Paul S. Bradley. "Efficient discovery of error-tolerant frequent itemsets in high dimensions." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.
[3] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases." *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.
[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. "Finding interesting rules from large sets of discovered association rules". *In CIKM'94*, pp. 401-408.
[5] J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. *In SIGMOD'95*, pp. 175-186
[6] Michael Steinbach and Vipin Kumar " Generalizing Notion of Confidence" *Knowledge and Information Systems*, 12:279-29 Jan 2007;
[7] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In VLDB'95, pp. 432-443.
[8] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In KDD'97, pp.67-73.
[9] Koh, Jia-Ling, and Pei-Wy Yo. "An efficient approach for mining fault-tolerant frequent patterns based on bit vector representations." Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2005.
[10] Dongre, Jugendra, Gend Lal Prajapati, and S. V. Tokekar. "The role of Apriori algorithm for finding the association rules in Data mining." Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on. IEEE, 2014.
[11] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data."Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.
[12] Jiawei Han, Micheline Kamber "Data Mining Concepts and Technigues" 3rd Edition, pages 227-245, 2006
[13] S.-S. Wang and S.-Y. Lee, "Mining Fault-Tolerant Frequent Patterns in Large Database,"in Proc. Of Workshop on Software Engineering and Database Systems, InternationalComputer Symposium, Taiwan, 2002
[14] Lin, Ying Chun, Cheng-Wei Wu, and Vincent S. Tseng. "Mining high utility itemsets in big data." Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2015. 649-661

[15] Christian Borgelt,Christian Braune1,Tobias Kotter and Sonja Grun "New Algorithms for Finding Approximate Frequent Item Sets"Soft Comput 16:903–917, 2012
[16] Steinbach, Michael, et al. "Generalizing the notion of support." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
[17] Pei, Jian, Anthony KH Tung, and Jiawei Han. "Fault-Tolerant Frequent Pattern Mining: Problems and Challenges." DMKD 1 (2001): 42.
[18] Jiawei Han, Michelin Kamber " Data Mining: Concepts and Techniques, 2nd Edition, 2006