International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com‖ Volume 2 Issue 3‖ March 2016 ‖ PP 46-51

A simple framework for calculating metrics of Fault Tolerant Itemsets

Morrel VL Nunsanga
Dept. of Information Technology
Mizoram University, Aizawl, Mizoram, India

Abstract- *The Association rule mining searches for interesting relationship among items in a given data set. There are still big challenges in finding out interesting relationship among items in large volume of data as these big volumes of data tend to contain errors or noises. In presence of such errors, the traditional approach for calculating support and confidence gives values which are at times meaningless. In this paper, the framework for computing the standard metrics i.e. confidence and support for Fault Tolerant Itemsets(FTI) is given, and in addition to the standard metrics, a new metric called, Interestingness is also introduced to add additional information to association rule for FTIs.*

Keywords – *Fault Tolerant Itemsets(FTI), Interestingness, confidence, support*

I. INTRODUCTION

Progress in database technology has generated rapid growth in the storage of information and data in large volume and these transaction data are ubiquitous. The frequent itemset problem is that of determining which items frequently occur together in transactions. But in a large volume of transaction data, there are big chances of occurrence of errors or noises due to different reasons such as human errors, or the items could not be purchased as it was out of stock, or due to minor defect in the machine, etc.In such case, the classical definition of frequent itemset where exact matching is needed may miss many useful patterns in the data set.

Defintion 1 :Fault Tolerant Itemset

An itemset $E \subseteq I$ is a fault tolerant itemset having error $\mathcal{E}$ 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- $\mathcal{E}$ of the items from E are present[1]

Under standard definition, an exact matching is required to have a transaction supports an itemset if it contains a 1 under each item in the itemset. An itemset is said to be frequent if the number of its supporting count exceeds the "support threshold," a user determined percentage of the total number of transactions.

Fault tolerant itemsets relax the requirement that every transaction supporting the itemset must contain every item. Instead, it is enough that each transaction contains most of the items in the specified itemset. At the same time, extracting fault tolerant frequent itemset may produce many unnecessary patterns and it is difficult to have the actual measure of the relation of the itemsets. i.e. the two existing measures of the itemsets 'confidence' and 'support' do not sometimes reveal the actual strength of the relation of the itemsets. For example, a particular item which is having very less occurrence in the transaction may be included in the frequent itemset count if the error in the set satisfies threshold value. In such situation, the confidence value may not reveal the actual relationship among the items in the frequent set.

An easy method for calculating confidence and support for the rule is provided by constructing boolean vectors and applying the required logical operations on the vectors. In addition to this new method, a new metric called 'Interestingness' has been introduced to give more intuitive meaning to the association rule. The measures of interestingness in the rule give the participation of each item in frequent set.

II. BACKGROUND AND RELATED WORKS

Lots of researches and works have been done on frequent itemsets problem, and lots of algorithms have been proposed already. Among these, the classical Apriori proposed by Agrawal, Rakesh, and Ramakrishnan Srikant [2] is one the most common and popular on this particular problems. The anti monotone property provides an efficient way to find frequent itemsets and is the main foundation of the Apriori. But this property no longer holds as the notion of errors is introduced into the definition of frequent itemsets. So, there are lots of challenges on error tolerant frequent itemsets; thorough researches and studies have to be done on this error frequent itemsets problems.

The error tolerant frequent set problems were first discussed by Yang, Cheng, Usama Fayyad, and Paul S. Bradley[1] who classified error tolerant into two types: weak error tolerant itemset and strong error tolerant itemset. They also proposed algorithms to find maximal weak ETIs and strong ETIs from datasets. Though the time complexity of the first algorithm, Exhaustive Growing Algorithm, is quite high, in practice, it is observe that the chance of missing ETIs is very low.

M.Steinbach and V. Kumar [3] discussed the problems in computing the confidence and the support of the association rule of ETIs. They provide a frame work for defining confidence that can be extended to ETIs and even to continuous data. But, even this framework does not see the position of the occurrence of the error. There are many other researches and studies on the error tolerant itemsets[4-14], but lot more challenges need to be explored. This research work is to provide a simple method to calculate the metrics of the pattern and to analyze the actual contribution of each item including errors or noises in the itemset

A. Traditional approach for calculating support and confidence

Definitin 2 : Support

The support for a set of items is the percentage of transaction that contains all of these items. The support for a particular association rule X => Y is the proportion of transactions in D that contain both X and Y.

Support (X => Y) = $P(X \bigcup Y)$

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

Definition 3: Confidence

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. In other words, confidence is the ratio (in percent) of the number of records that contain X $\bigcup$ Y to the number of records that contain X.

Confidence (X=>Y) = $P(Y|X)$

$$= \frac{\text{Sup } (X \bigcup Y)}{\text{Sup } (X)}$$

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

B. Traditional Approach to Confidence for Fault Tolerant Itemsets

For binary transaction data, the traditional approach to confidence for two itemsets, *X* and *Y* is based on support, i.e., $Conf(X => Y) = Sup(X \bigcup Y) / Sup(X)$. Now, let us calculate the confidence for (X=>Y) in the following data show in the table 1. Let X = {i1, i2, i3}, Y={i4, i5}, and $\mathcal{E}$ =0.4, k=50%, then the given sets are fault tolerant itemsets .

Table 1: Sample data set with few errors

	i1	**i2**	**i3**	**i4**	**I5**
t1	1	1	0	1	1
t2	1	0	0	1	0
t3	1	1	1	1	1
t4	1	0	1	0	0
t5	1	1	0	1	1
t6	1	0	1	1	1

From the above table, $Sup(X \bigcup Y) = 4$, $Sup(X) = 5$

$Conf(X => Y) = Sup(X \bigcup Y) / Sup(X) = 4/5 = 0.8$ or 80 %

$Sup(X=>y) = Sup(X \bigcup Y) / T = 4 / 6 = 0.67$ or 67 %

III. PROBLEMSSTATEMENT

In fault tolerant itemsets, the traditional approach may not be suitable for calculating the metrics of the itemsets at times. Two cases have been highlighted below where the traditional approach seems inappropriate:

Case 1: The support and confidence of the frequent itemset reveal the strength of the relationship of the itemset. But in Boolean based transaction in Fault Tolerant Itemsets, the positions of the assigned values (0 or 1) and the

threshold values greatly affect the result of the confidence . Sometimes, the ratio in traditional approach becomes meaningless i.e confidence of Fault Tolerant Itemsets does not reveal any relationship strength.

Consider the following table

Table 2 : Sample data set to illustrate problems on confidence value for FTIs using traditional approach

	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i8**	**i10**
t1	1	1	1	1	1	0	0	0	0	0
t2	1	0	1	1	1	0	0	0	0	0
t3	1	1	0	1	1	0	0	0	0	0
t4	1	1	1	0	1	0	0	0	0	0
t5	0	1	1	1	1	0	0	0	0	0
t6	0	0	0	0	0	0	1	1	1	1
t7	0	0	0	0	0	1	1	1	1	0
t8	0	0	0	0	0	1	1	1	0	1
t9	0	0	0	0	0	1	1	1	0	1
t10	0	0	0	0	0	1	0	1	1	1

Suppose $\mathcal{E}$ =0.6, *k*=50% (i.e. at least a transaction must contain 4/10 items and half of the transaction support the patter. If X = {i1, i2, i3, i4, i5} and Y = {i6, i7, i8, i9, i10}, then both X and Y are FTIs. Using traditional approach,

$$Conf(X => Y) = Sup(X \bigcup Y) / Sup(X) = 10/5 = 2$$

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

1) the value in any relationship should not be greater than 1 to reveal the actual strength of the association rule. But the confidence value obtained here is more than 1 which is really odd and does not reveal the actual strength of the relationship of the rule. This situation happens here because the support count of union of X and Y is more than the support count of X alone which is not supposed to happen.
2) looking at the itemsets of both sides of the rule(X and Y) the confidence value is expected to be zero (0) because the FTI patter in X never co-occurs with the FTI patter in Y,

Thus, the traditional notion of confidence does not seem appropriate for FTIs.

Case 2: When a number of 1s in a column is too less:

Consider the following table which has a column with very less 1s

Table 3: Data set where there are too few 1s in a column

	i1	**I2**	**i3**	**i4**	**i5**	**i6**
t1	1	1	1	1	0	0
t2	1	1	1	1	0	0
t3	1	1	1	1	0	0
t4	1	1	1	1	0	0
t5	1	1	1	1	1	0
t6	0	0	0	0	1	1

From the above table the itemset {i1, i2, i3, i4, i5} can be regarded as FTIs with *k*=2/6 (33.3%) and $\mathcal{E}$ = 20 % (0.2). But actually i5 has very less number of 1s and can be neglected and does not have much importance in practical.

IV. PROPOSED APPROACH

A. Framework for calculating Confidence and support :

To calculate the confidence of the rule (X=>Y), where '*T*' is the total number of transactions, '$\mathcal{E}$' is error tolerant threshold and '*k*'support threshold:

- The left side itemset X and right side itemset Y of the rule are checked against the error tolerant value

(ε) in each transaction

- Construct vector set $V_{xt} = \{V_{xt1}, V_{xt2}, V_{xt3}.....V_{xtn}\}$ where $V_{xti} = 0$ if X satisfies ε for t^{th} raw,else $V_{xti} = 1$
- Construct vector set $V_{yt} = \{V_{yt1}, V_{yt2}, V_{yt3}.....V_{ytn}\}$ where $V_{yti} = 0$ if Y satisfies ε for t^{th} raw,else $V_{yti} = 1$
- AND the value of V_{xt} and V_{yt}; increment the support count of the rule S_{xy} whenever the ANDing result is 1
- Repeat the procedure for all transactions. The left side itemset support count S_x is also recorded.

Confidence: To compute the confidence of the rule, the final S_{xy} value is divided by the support count of the left side itemset S_x

Mathematically, it is shown below:

$$\text{Conf}(X=>Y) = S_{xy}/S_x = \frac{\sum_{t=1}^{T} Vxt \wedge Vyt}{\sum_{t=1}^{T} Vxt}$$

where V_{xt} = 1 if X satisfies ε for t^{th} row else = 0

V_{yt} = 1 if Y satisfies ε for t^{th} row else = 0

Support: To compute the support of the rule, the final S_{xy}value is divided by the total number of transactions '*T*'

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

B. Introducing Interestingness :

In this paper a new metric called *Interestingness* consequently *Interestingness Threshold* (ρ)has been introduced. The Interestingness metric shows the fraction of the presence of each item of the item set in the supported transactions. For instance, an *Interestingness* of a particular item of 60% means that percentage of the occurrence of that particular item in the supported transactions of the itemset is 60%

The items which are having *Interestingness* value less than the *Interestingness Threshold* (ρ) will be explicitly displayed. The formal definition of *Interestingness* is given below

Definition 3: Interestingness

Interestingness *of an item in a given itemset is a fraction of the number of presence of the item in the transaction of the ETI supported rows.*

It can be calculated as below:

$$\text{Interestingness}(I_i) = \sum_{t=1}^{R} (Ii_t) \Big/ \sum_{t=1}^{R} (Xt)$$

where I_{it} is the i^{th} item of the *I* itemsets in the FTI supported t^{th} row , *Xt*= 1 when the itemset (X) satisfies the error tolerant threshold. R is the number of the FTI supported rows.

V. RESULT

The proposed approached was applied to calculate confidence and support of different scenario of datasets and the same result with the traditional approach were obtained in many cases. It can further be used to solve the problems explained in previous section and the value obtained using the proposed approach give much more intuitive meaning for the relationship among the items. How the proposed approach provides solution to the two sample cases is demonstrated below

Solution to the case 1: Consider the table shown in the previous chapter(Table 2) where ε =0.6, k = 50%
let ρ =50%

We have , $\text{Confi}(X->Y) = \frac{\sum_{t=1}^{T} Vxt \wedge Vyt}{\sum_{t=1}^{T} Vxt}$

V_{xt} = [1,1,1,1,1,0,0,0,0,0], therefore, $S_x = 4$
V_{yt} = [0,0,0,0,0,1,1,1,1,1]
$V_{xt} \Lambda V_{yt}$ = [0,0,0,0,0,0,0,0,0,0], therefore, Sxy = 0 (ie. Sum of $V_{xt} \Lambda V_{xy}$)
Conf(X=>Y) = S_{xy}/S_x
= 0/4 = 0

The confidence value 0 here is much more intuitive than the previous result computed with traditional approach. Interestingness is not needed to be shown since confidence is zero.

Solution to Case 2: In case of less occurrence of an item in a supported transactionas given in the sample dataset given in table 2 where $\varepsilon = 0.2$, k=0.5, ρ =0.5

We have, Confi(X->Y) = $\dfrac{\sum_{t=1}^{T} Vxt \wedge Vyt}{\sum_{t=1}^{T} Vxt}$

= 3/5 = 0.6 = 60 %

Here, Interestingness (I_5) = $\sum_{t=1}^{R} (I_{5t}) \; / \; \sum_{t=1}^{R} (Xt)$

= 1/5 = 0.2 = 20 %

All the other items have *interestingness* value more than the threshold
So, we can write for the association rule: {i1, i2, i3, i4, i5} => {i6, i7, i8}

Support = 4/8 = 0.5 = 50 %
Confidence = 3/5 = 0.6 = 60 %
With Interestingness (I_5) = 1/5 = 0.2 = 20 %

This information with additional *Interestingness* would definitely reveal the actual relationship among the items in the itemset

Analysis of the new metrics

- The new metric 'Interestingness' does not show strength of the rule as a whole rather it shows the strength of the individual item in the pattern where it occurs.
- The 'interestingness' of an item should be calculated with respect to the FTI supported rows. i.e. interestingness values are calculated from the transactions where both sides of the rule satisfy support threshold.
- The 'Interestingness' adds additional information which could be very useful for retailer in making decision for inventory management, developing marketing strategies, etc.
- It does not have effect on the confidence value of the association rule.
- The interestingness of items in a conventional frequent itemset are 100% each

The experimental results show that a meaningful value always been observed with the proposed approach and the newly introduced metric called,*Interestingness* provides additional information for the relationships of each item in the rule.

VI. CONCLUSION

A boolean vector based framework for calculating confidence and the support of the association rule for Fault Tolerant Itemsets have been presented in this paper. In some cases of Fault Tolerant Itemsets, the calculated value of the standard metrics i.e. support and confidence sometimes do not reveal the actual information of the association rule, especially when errors are not distributed equally in the item set. It has been demonstrated that how this framework solves such problems and generated much more intuitive value of support and confidence. In addition to this standard metrics, the new metric, called 'Interestingness' has been introduced to add additional information on each item of the itemset. This additional information on the association rule improves the usefulness of the rule in real life application.

Working with large volume of data in presence of noises is still a big challenge. This framework is in on binary based transactions and would be interesting if the work is extended to explore a continuous-value domain.

VII. REFERENCES

[1]. 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.

[2]. Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[3]. Michael Steinbach and Vipin Kumar " Generalizing Notion of Confidence"Knowledge and Information Systems, 12:279-29 Jan 2007;

[4]. 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.

[5]. 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

[6]. 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.

[7]. Jiawei Han, Micheline Kamber "Data Mining Concepts and Technigues" 3rd Edition, pages 227-245, 2006

[8]. J. Liu, S. Paulsen, Wei Wang, Andrew Nobel, Jan Prins, "Mining Approximate Itemsets From Noisy Data"TechnicalReport(TR05-015) of Department of Compuater Science, UNC-Chapel Hill, Jun 2005.

[9]. Christian Borgeltand Tobias Kotter "Mining Fault tolerant Item sets using Subset Size occurrence Distribution" In proceeding of Advances in Intelligent Data Analysis X - 10th International Symposium, IDA 2011, Porto, Portugal, October 29-31, 2011.

[10]. Christian Borgelt,Christian Braune1,Tobias Kotter and Sonja Grun "New Algorithms for Findin Approximate Frequent Item Sets"Soft Comput 16:903–917, 2012

[11]. 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.

[12]. Pei, Jian, Anthony KH Tung, and Jiawei Han. "Fault-Tolerant Frequent Pattern Mining: Problems and Challenges." DMKD 1 (2001): 42.

[13]. 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.

[14]. Lin, Ying Chun, Cheng-Wei Wu, and Vincent S. Tseng. "Mining high utility itemsets in bi data." Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2015 649-661.