International Journal of Pure and Applied Mathematics
Volume 117 No. 21 2017, 799-810
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Performance Measure of Exceptional Frequent Pattern Mining using Optimized Sampling from Large Database

[1]Pragnyaban Mishra, [2]Sudhanshu Shekhar Bisoyi and [3]Sarojnanda Mishra

[1]Department of Computer Science and Engineering,

KL University, Vijayawada, AP, India.

pragnyaban@gmail.com

[2]Department of Computer Science and Engineering,

KL University, Vijayawada, AP, India.

sudhanshu.bisoyi@gmail.com

[3]Department of Computer Science and Engineering,

IGIT, Sarang, Dhenkanal, Odisha, India.

sarose.mishra@gmail.com

Abstract

The process of identifying exceptional pattern association from a large database is a challenging task. These are generally uncertain or exceptionally occurring patterns with a very high frequency. In the last two decades many frequent pattern mining algorithms have been proposed to extract knowledge from these frequent patterns. Most of them require multiple scan over the entire database. In order to do the transaction reduction many sampling based frequent pattern mining approaches have been proposed for increasing the efficiency in the mining or pattern associations. In this paper first, we have extracted all frequent itemset from the initial sample size. In the second, an optimized technique has been proposed to select the appropriate sample size. Then we have analyzed the optimized sampling based apriori approach over the traditional approach while generating the required frequent pattern or association rules. The experimental result analysis has been shown that our optimized approach is better than the traditional apriori.

Key Words: Exceptional pattern, frequent pattern, association rule, apriori, sampling.

1. Introduction

With increasing development in information and communication technology a huge amount of data are used to be generated from various sources. These data need to be collected and stored for knowledge discovery or mining. These collected data are having different structure, different geographical property or characteristics which are growing exponentially in a diversified structure or format. Many data mining techniques and algorithms are available for extracting useful information by eliminating the unwanted patterns from the database. Data mining is a fundamental and crucial approach for extracting useful and understandable information or knowledge from the huge amount of data for any organizational Decision Support System (DSS). It makes use of methods, algorithmic approach and techniques from various disciplines in order to mine understandable and valuable information from a massively collected database. It is really a challenging task to make use of appropriate algorithms or optimized technique to get a desirable solution while performing data mining or data analytics. Among all these technique association rule mining is the most widely used approach in various fields like market basket analysis, recommendation system, health care, finance, security etc. Association rule mining aims to discover useful association among the set of frequently occurring items, patterns, events or objects. In general the data mining or analysis tasks can be classified into two types: Descriptive analysis and Predictive analysis. Descriptive data analysis is to determine interesting properties and to summarize the data, whereas predictive data analysis focuses in designing a model to identify knowledge. Clustering, Association Rule Mining and Sequential Pattern mining are commonly used in descriptive data analysis.

In a transactional database it has been observed that many patterns or items are occurring rarely. Sometime these rare items occurrence leads to very high frequency and carry vital information. These rarely occurring frequent itemset may be considered as the exceptional itemset in the database [1,5]. To extract associated knowledge these exceptional patterns or itemset need to be mined from the huge amount of dataset. During the extraction of useful information for the exceptional patterns in large database the efficiency is considerably affected by system resources like processing and IO transfer. In order to improve the processing efficiency sampling is used. The sampling technique is used to reduce the dataset size and help in estimating the overall knowledge of the whole dataset. Though sampling is used as a data reduction technique while mining the knowledge using association rule, it may not guarantee in accuracy. There is a chance of deviation in accuracy because of the varying size in the dataset of both sample and population. It may leads to generation of unwanted association rules and wrong DSS.

In this paper we have taken exceptional transactional database and applied an optimized sampling technique which is a combination of stratified and

progressive sampling. The association rules are generated from the initial sample with support and confidence. We compare all the generated rules of sample with respect to sub population known as stratum. Testing is applied to check accuracy of the rules, if the accuracy is below the acceptance level then sample size has to be optimized .In the remaining part of our paper in section 2 we have explained the background study on sampling based association rule mining, section 3 represent our proposed methodology, in section 4 the Experimental study and section 5describes the Conclusion.

2. Background Study

Many researches have been done on sampling based association rule mining. At first a single scan algorithm for mining frequent patterns has been proposed by Toivonen in the year 1996. He has proposed a simple technique in which a random sample of the large database is used for generating frequent itemsets and validated in a single database scan. His approach not only counts the frequent itemsets during the scan but also compute the negative border. If the frequent itemsets generated during the scan are in the negative boarder then one more scan has to be performed for identifying whether any subset of these items are frequent or not. In order to reduce the failure, a lower support parameter has been used and compared with the minimum support of the frequent itemset. The results reported only if they pass the original *minSup* threshold. In the empirical study of sampling based association rule the Chernoff bounds can be used to determine the sample sizes to achieve a desired degree of accuracy which is independent of the size of database.

A two-phase sampling-based algorithm, FAST (Finding Associations from Sampled Transactions) for mining association rules in large database have been proposed in [3]. In Phase I, a substantial initial sample of transactions was collected and then used to measure the support of each individual item in the large database quickly and accurately. In Phase II, these transactions are eliminated so that representative transactions can be chosen from the initial sample. This helps in collecting a small and accurate sample that more correctly reflects the statistical characteristics of the entire database.

In the recent years many studies have also been conducted on exceptional or uncertain pattern mining based on stratified and progressive sampling for the effective association rule generation [2, 4, 5, 7, 9]. It has been clearly studied that the database contains many uncertain or exceptional items in a particular time period. The exceptional data is not a regular data, it is mostly uncertain. Consider the demand of the sweets, this increases exceptionally during the festive season, as a result transactions dataset contain high amount of sweet items during that season where as it is less on other season. We cannot ignore those data items for future study and analysis.

Probability is a good measure to choose items inside the sample [2] e.g. for every transaction t and every item i in a transactional database, an independent

random number r is generated $0 \le r \le 1$and we compare it with the probability p associated with the item i. If $p \ge r$ then item i will appear in the currently sampled transaction. Philip S. et. al. in [5] have represented a survey on uncertain data, which is defined as a collection of instances of data in an incomplete form or probabilistic quantification. The probabilistic information in the database is a finite probability space whose outcomes are all possible database instances consistent with a given schema. This can be represented as the pair of (X, p) where X is a finite set of possible database instances consistent with given schema and $p(I)$ is the probability associated with any instances $I \in X$. The existential uncertainty is a probabilistic measure [8]. It is defined as a tuple that may or may not exist in the database, but affect the probability of the presence or absence of another tuple in the database. It is a useful application of data mining where an user understand the occurrence and significance of items in transaction database. The existential probability $p(x_i, t_j)$ is used to find out the possibility of item x_i in the transaction of t_j. It has been clearly studied that existential probability is used in exceptional pattern mining [8]. The following definition describes about the exceptional support and uncertain data:

Def 01: Let Item be a set of m domain items. Each item y_i in a transaction $t_j = \{y_1, y_{2,}.....y_h\} \subseteq$ Item in a probabilistic set of uncertain data is associated with an existential probability denoted $p(y_i, t_j)$ with value $0 \le p(y_i, t_j) \le 1$ where $p(y_i, t_j)$ is represents likelihood of the presence of y_i in t_j.

Def 02: Let Item be a set of m domain items. Then, the expected support denoted as $\exp Sup(X, t_j)$ where $X = \{x_1, x_2, \ldots, x_k\} \subseteq item$ in a transaction $t_j = \{y_1, y_2, \ldots, y_r, \ldots, y_h\}$ where $x_k = y_r$ can be computed as the product of the existential probability $p(y_i, t_j)$ of every independent item y_i within the itemset X i.e.,

$$\exp Sup(X) = \sum_{j=1}^{n} \exp Sup(X, t_j).$$

where $\sum_{j=1}^{n} \exp Sup(X, t_j)$ can be computed as the product of the existential probability $p(y_i, t_j)$ of every independent item x_i within the itemset $X = \{x_1, \ldots, x_k\}$.

Pattern sampling has been proposed as a potential solution to the pattern explosion. Instead of enumerating all patterns that satisfy the constraints, individual patterns are sampled proportional to a given quality measure. Several sampling algorithms have been proposed, but each of them has its limitations

when it comes to (1) flexibility in terms of quality measures and constraints that can be used, and/or (2) guarantees with respect to sampling accuracy [6]. An important issue in any estimation is that the estimator can take care of all salient features of the population. If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample and in turn, the sample mean will serve as a good estimator of population mean. In order to increase the precision of an estimator, we have to use a sampling scheme which reduces the heterogeneity in the population using stratified sampling. In [10] stratified sampling techniques have applied in deep web and found the result as a tree model to capture the relation between input and output attributes of the deep web data source, a recursive statistical method has been used to reduce the integrated cost metric that combines estimated variance, sampling cost and an optimized sample allocation method that takes into account both the estimation error and the sampling costs.

Application and Uses of Stratified Sampling

In order to find the average weight of chicken in a farm there will be subpopulation from the population as per their age:-

Stratum1: $15\ days \le age < 25 days.$

Stratum 2: $25\ days \le age < 35 days.$

Stratum 3: $35\ days \le age < 45 days.$

Stratum 4: $45\ days \le age < 55 days.$

Now draw the samples by SRS from each of the strata 1, 2, 3 and 4. All the drawn samples combined together will constitute the final stratified sample for further analysis. It has been observed that the item present inside a stratified sample is homogeneous in nature where as among the samples it is heterogamous. In stratified random sampling, the population of size N is partitioned into strata and a sample is selected by simple random sampling within each stratum [2]. Given a total sample size n if the strata differ in size, different allocation could be used to maintain a steady sampling fraction throughout the population. If stratum k has N_k units, the sample size allocated to it would be

$$n_k = \frac{n}{N} N_k \qquad \text{Eq(1)}$$

When the sample is not adequate with the population we may increase the size of the sample using progressive sampling. Hence our approach is leading towards progressive stratified sampling. In [7] the uses of progressive sampling in mine k item set has been explained and extended the analysis of progressive sampling over real life datasets.

3. Proposed Methodology

It has been very clearly studied that the exceptional patterns are not occurring very frequently in the regular transactional dataset. Their occurrence is very high in a particular time period and then these patterns or items become irregular in nature. In general it is very difficult to find the frequently occurring exceptional itemset form a very large collection of regular transaction or market basket database. The traditional association rule mining algorithm for the exceptionally occurring high frequent itemset leads to the difficulty in generating:

- Large number of unwanted frequent patterns.
- Large number of unwanted association rules.
- Large amount of memory space to store these unwanted frequent pattern and association rules.

The primary concern with our proposed method is to generate high frequent itemset or patterns along with the association rules from the exceptional itemset in the transactional dataset. The sampling approach has been used to discover the association rules by choosing appropriate sample size. The optimal sample size need to be collected such that the accuracy will be more in concerned to the original dataset known as *SampApriori* algorithm. We have used the following steps for mining exceptional frequent pattern and association rules by considering an optimal sample size.

- Select a sample from strata of sample population.
- Use the apriori algorithm for generating the entire frequent pattern.
- Compute the existential probability support and determine the exceptional items.
- Compute the accuracy for all generated rules.
- If the accuracy is not obtained then an optimized sample size has to be selected and the accuracy needs to be computed again for population of patterns.

3.1. Algorithm: SampApriori

Input: Sub population database D_k, Size of initial sample in %, increment of sample size in %, minsupp, minconf. Here k represents the k^{th} subpopulation of the population α is minimum support and β is minimum confidence.

Output: Exceptional Rules.

Step-1: *Initialize,* $k \leftarrow 1 \ and \ c \leftarrow Total\_Strata$

Step-2: *Initialize the sample* S_{ik} *of size* n_k *from the sub population Database* D_k *of size* d_k at random

Step-3: *Generate frequent itemsn with upper and lower support bound by using apriori algorithm for sample and population.*

$$f\_Items_{samp} = getFreqItems(S_{ik}, \alpha)$$

$$f\_Items_{popul} = getFreqItems(D_k, \alpha)$$

Step-4: *Generate all the association rules* R_{samp} *from these frequent items* f_Items_{samp}.

Step-5: *Generate all the association rule* R_{popul} *from the remaining items in the database* f_Items_{popul}

Step-6: *Select the rules from* R_{samp} *and* R_{popul} *where* $\alpha < \{R_{samp}, R_{popul}\}$ *and* $\beta > \{R_{samp}, R_{popul}\}$

Step-7: *Compute the accuracy between the selected sample itemsets and total items in the database. If the accuracy is achieved then the sample size is optimized. Otherwise, the sample size 'n' is increased and steps 3-7 are performed progressively until an optimal sample size is achieved.*

Step-8: if $(k \le c)$ *then* $k = k+1$ *and goto step-2*

Step-9: Write the optimized sample sizes with the exceptional frequent items.

The frequent itemset can be generated with the help of apriori algorithm. The pseudocode of apriori is given as follows:

3.2. Algorithm-apriori

Input: Items in EDB and α

Output: set of frequent itemset

(1) $L_1 = getFreqItemset_1(EDB)$ where EDB is the exceptional database
(2) for $(k = 2; L_{k-1} \neq \varnothing; k++)$ {
(3) $C_k = apprioriGen(L_{k-1}, \alpha)$ where α is min support
(4) for each transaction $t \in EDB$ { //scan EDB for counts
(5) $C_t = subset(C_k, t)$; // get the subsets of t that are candidates
(6) for each candidate $c \in C_t$
(7) c.cnt++;
(8) }
(9) $L_k = \{c \in C_k \mid c.cnt \ge \alpha\}$;
(10) }
(11) return $\bigcup_k L_k$

4. Experiment

In this section we have described the experimental result and analysis of optimized sampling technique while mining association rule from large database. The sample size has been chosen very carefully so that the accuracy in

DSS can be increased. In our research work we also have performed a comparative study between traditional apriori algorithm with stratified sampling and our optimized sampling approach. Different measures have been considered during the performance analysis. Laplace, lift and conviction are the good measures of association rule if dataset is small using Eq(2), Eq(3), Eq(4) .

$$conf(X \to Y) = \frac{\sup(X \cup Y)}{\sup(X)} \qquad \text{Eq(2)}$$

$$lift(X \to Y) = \frac{conf(X \to Y)}{\sup(X)} \qquad \text{Eq(3)}$$

$$lapl(X \to Y) = \frac{\sup(X \cup Y)+1}{\sup(X)+2} \qquad \text{Eq(4)}$$

In our research work we have considered different transactional dataset and applied an optimized sampling technique to the input data set. We have measured the accuracy and errors as per given below Eq(5), Eq(6) and Eq(7). If the sample is not adequacy with the population output then we may increase the size of sample using the progressive sample algorithm.

4.1 Measurement of Accuracy and Errors

$\Omega(D)$: All frequent itemsets generated from the original database.

$\Omega(S)$: All frequent itemsets generated from the sample data.

Let an itemset present in $\Omega(S)$ but not in $\Omega(D)$ then it is a false positive. Other than this it is called as false negative.

$|\Omega(S) - \Omega(D)|$: The number of false positives.

$|\Omega(D) - \Omega(S)|$: The number of false negatives.

Measure 01: The accuracy of the sampling is measured as:

$$accuracy = 1 - \frac{|\Omega(D) - \Omega(S)| + |\Omega(S) - \Omega(D)|}{|\Omega(D)| + |\Omega(S)|} \qquad \text{Eq(5)}$$

Besides this following two measures are used to distinguish difference between sample and population.

Measure 02: Quantification of the sampling error

$$f(p) = \frac{|\Omega(S) - \Omega(D)|}{\Omega(S)} \qquad \text{Eq(6)}$$

Measure 03: To represents the proportion of the false frequent itemsets in a sample, and the proportion of the frequent itemsets in the original dataset that is

missing in a sample

$$f(n) = \frac{|\Omega(D) - \Omega(S)|}{\Omega(D)} \qquad \text{Eq(7)}$$

One large population of sample has been considered while carrying out the experiments analysis. Multiple stratums have been chosen from the sample population and a sample has been taken at random from the strata. The frequent items and the corresponding association rules have been generated from that sample. If the generated frequent itemset satisfy the lower support and confidence boundary in that sample then these are chosen for organizational DSS. If not satisfied then the sample size is increased at random to achieve the support and confidence boundary. The sample based association rule mining has been implemented using Java programming language on Linux platform. Figure-1 describes the accuracy of the selected sample in the strata of the sample population. It has been shown that accuracy of the selected samples of all the strata from the sample population is almost similar or slightly changing. The Figure-2 describes the performance of apriori with simple random sampling and optimized sampling with varying support while computing the frequent pattern and association rule. Our optimized sampling based apriori algorithm shows better performance as compared to random sampling based apriori algorithm. Figure-3 describes the accuracy by considering different support in the selected sample of the population and the Sample apriori has shown better accuracy as compared to random sampling based apriori algorithm.

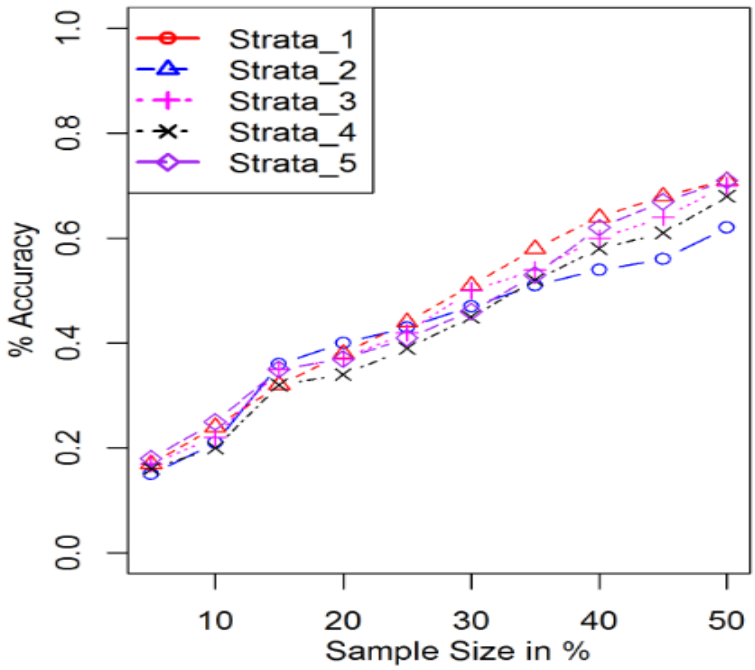

Fig. 1: Sample Size vs Accuracy

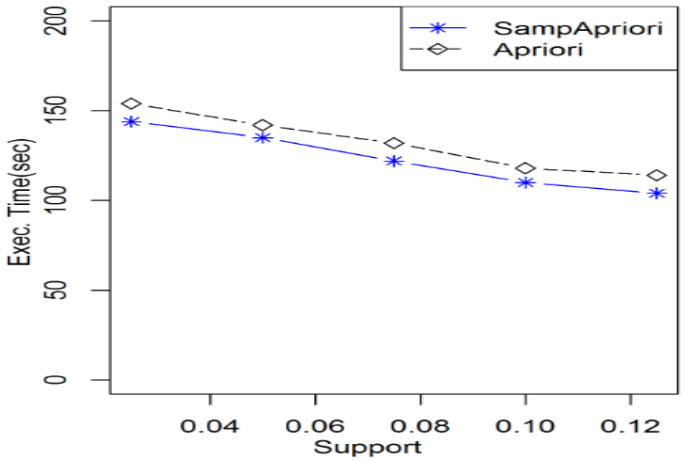

Fig. 2 Support vs Execution Time(in sec)

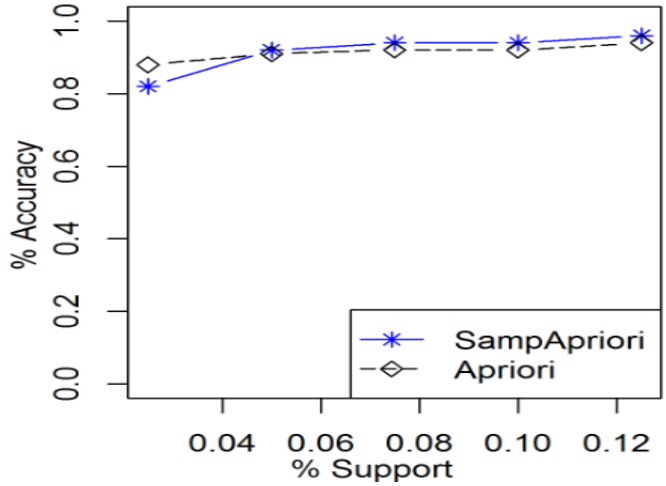

Fig. 3: Support vs Accuracy

5. Conclusion

In this paper we have performed an extensive analysis of selecting an optimized sample such that the accuracy of association rule mining can be increased. We have focused on the mining of exceptional pattern from a large amount of dataset with the help of optimized sampling based apriori algorithm. In our approach an initial sample has to be chosen at random from the given set of population. After selecting the initial sample we have checked the accuracy while generating the frequent pattern. In order to get the optimized sample we have applied a hybrid sampling approach. The basic performance comparison with respect to accuracy and IO overhead has been performed between the traditional apriori and our optimized sampling based apriori algorithm. As the sample size we have determined is optimized, it produces more accurate result. The experimental result has also ensured that our optimized approach produce accurate result while mining the association rule mining.

References

[1] Philip S. Yu, Charu C. Aggarwal, A survey of uncertain data algorithms and applications, IEEE Transactions on Knowledge Data Engineering 21 (2008), 609-623.

[2] Yanrong Li, Raj P. Gopalan, Stratified Sampling for Association Rules Mining, Springer US, Boston, MA (2005), 79-88.

[3] Bin Chen, Peter Haas, Peter Scheuermann, A new two-phase sampling based algorithm for discovering association rules, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002), 462-468.

[4] Toon Calders, Calin Garboni, Bart Goethals, Efficient Pattern Mining of Uncertain Data with Sampling, Springer Berlin Heidelberg, Berlin, Heidelberg (2010), 480-487.

[5] Qun Yu, Ke-Ming Tang, Shi-Xi Tang, Xin Lv, Uncertain Frequent Itemsets Mining Algorithm on Data Streams with Constraints, Springer International Publishing, Cham (2016), 192-201.

[6] Vladimir Dzyuba, Matthijs van Leeuwen, Luc De Raedt. Flexible constrained sampling with guarantees for pattern mining, Data Mining and Knowledge Discovery 31(5) (2017), 1266-1293.

[7] Andrea Pietracaprina, Matteo Riondato, Eli Upfal, Fabio Vandin, Mining top-k frequent itemsets through progressive sampling, Data Min. Knowl. Discov. 21(2) (2010), 310-326.

[8] Alfredo Cuzzocrea, Carson K. Leung, Richard Kyle MacKinnon, Approximation to expected support of frequent itemsets in mining probabilistic sets of uncertain data, Procedia Computer Science, 60(Supplement C) (2015), 613-622.

[9] Tom Au, Meei-Ling Ivy Chin, Guangqin Ma, Mining Rare Events Data by Sampling and Boosting: A Case Study, Springer Berlin Heidelberg, Berlin, Heidelberg (2010), 373-379.

[10] Tantan Liu, Fan Wang, Gagan Agrawal, Stratified sampling for data mining on the deep web, Frontiers of Computer Science, 6(2) (2012), 179-196.