ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Discovering Optimal Data Proficiency for Weighted Item Set Using Rank Pruning Model Based On Apriori Algorithm

B.Shanthini[1], G.Mathurambigai[2]

Research Scholar, Department of Computer Science, Swami Vivekananda Arts and Science College, Thiruvalluvar University, India[1]

Head, Department of Computer Science, Swami Vivekananda Arts and Science College, Thiruvalluvar University, India[2]

ABSTRACT: Outlier detection in high dimensional data presents various challenges resulting from the curse of dimensionality. It can provide insight into how some points appear very infrequently in k-NN lists of other points, and explain the connection between AntiHub, outliers, and existing unsupervised outlier detection methods. Even though they focus on detecting outliers in high dimensional data becomes less performance with high computational time. This article addresses the discovery of infrequent and weighted list items that is the infrequent weighted itemset, from dataset. To improve the efficiency, the identification of infrequent weighted items in transactions can be done first, for detecting the outliers easily. The proposed work contributes the techniques that may concentrates on decision making system that supports domain experts targeted actions based on the characteristics of the discovered weighted itemset and it will improve the efficiency of data mining tasks.

KEYWORDS: Outlier Detection, k-Nearest Neighbor, Apriori FP Mining.

I. INTRODUCTION

In this article, data mining approaches is to find frequent itemset from a dataset and derive association rules. Finding frequent itemset with frequency larger than or equal to a user specified minimum support is not trivial because of its combinatorial explosion. Once frequent itemset are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Here proposed a novel Efficient Rank Pruning algorithm based on the Apriori Algorithm for frequency and infrequency item set mining is an exploratory data mining technique widely used for discovering valuable correlations among data. To improve the performance efficiency, identification of infrequent weighted items in transactions are to be created and it would be more natural presumably more efficient, to work with meaningful dataset. A weighted frequency of occurrence of an item set in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. In particular focus our attention on two different frequent and infrequent item set support measures,

1. The frequent and infrequent-support-min measure, which relies on a minimum cost function, i.e., the occurrence of an item set in a given transaction is weighted by the weight of its least interesting item.
2. The frequent and infrequent-support-max measure, which relies on a maximum cost function, i.e., the occurrence of an item set in a given transaction is weighted by the weight of the most interesting item.

II. OBJECTIVE

The main objective of this article is outlier detection based on the infrequent itemset finding and to produce the better results when compared to other outlier detection algorithm which are previously used in the outlier detection mining. In this existing approach k-NN concept is used to determine the nearest neighbor list and number of outlier objects in the list. This is not efficient because of uncertainty of loss function.

To overcome this problem Apriori is used to improve the frequent or infrequent itemset mining process.

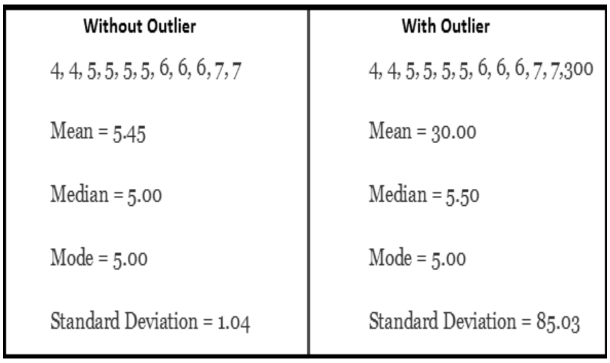
- To find infrequent weighted items in transaction dataset.
- To improve the efficiency of identification infrequent weighted items in transactions.

III. OUTLIER ANALYSIS

Hawkins formally defined the concept of an outlier as an outlier is and observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers.

Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process. The output of an outlier detection algorithm can be one of two types,

- Most outlier detection algorithm output a score about the level of outlierness of a data point.
- A second kind of output is a binary label indicating whether a data point is an outlier or not.



Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Fig.1 Outlier Data Vs. Original Data

IV. OUTLIER DETECTION

Detection of outliers in data defined as finding patterns in data that do not conform to normal behavior or data that do not conformed to expected behavior, such a data are called as outliers, anomalies, exceptions. Anomaly and Outlier have similar meaning. The analysts have strong interest in outliers because they may represent critical and actionable information in various domains, such as intrusion detection, fraud detection, and medical and health diagnosis.

Outlier detection and analysis are very useful for fraud detection, customized marketing, medical analysis, and many other tasks. Computer based outlier analysis methods typically follow a statistical distribution based approach, a distance based approach, a density based local outlier detection approach, or a deviation based approach. Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers. Outliers can be caused by measurement or execution error.

V. RELATED WORK: K-NEAREST NEIGHBOR ALGORITHM

Unsupervised outlier detection is done in a raw data collected from system. To identify unsupervised anomalies in high dimensional data is more complex. Therefore, the main objective of this thesis is to propose the unsupervised anomaly detection in high dimensional data. Anomaly detection in high dimensional data exhibits that as dimensionality increases there exists hubs and antihubs. Hubs are points that frequently occur in k nearest neighbor lists. Antihubs are points that infrequently occur in K-NN lists.

Outlier detection using AntiHub method is reformulated as Antihub^2 to refine the outlier scores of a point produced by the AntiHub method by considering Nk scores of the neighbors of x in addition to Nk(x) itself. Discrimination of outlier scores produced by Antihub^2 acquires longer period of time with larger number of iterations. Therefore Recursive AntiHub^2 method was introduced to improve the computational complexity of discriminating the outlier scores with reduced number of iterations to detect the more prominent outlier in high dimensional data. Such methods rely on a distance or similarity measure to find the neighbors, with Euclidean distance being the most popular option. Variants of neighbour based methods include defining the outlier score of a point as the distance to its k-th nearest neighbour.

The main objective is to increase the speed of computation and check the performance accuracy in discriminating the outlier scores with a threshold value. The computation speed is increased by reducing the number of iterations filtering the search area using binary search. Recursive AntiHub^2 method is introduced to do this process.

VI. FREQUENT PATTERN ITEMSET MINING BY APRIORI

The Apriori algorithm was proposed by Agarwal and Srikant in 1994. Apriori is designed to operate on datasets containing transactions. Apriori uses Breadth-First-Search and a hash tree structure to count candidate itemset efficiently.

1. Two novel qualities measures proposed to drive the infrequency data mining process. Infrequent item sets that do not contain any infrequent subsets have been proposed.
2. Experiments performed on both synthetic and real life data sets, show efficiency and effectiveness of the proposed approach.
3. In particular they show the characteristics and usefulness of the item sets discovered from data coming from benchmarking and real. To reduce the computational time the authors introduce the residual trees.
4. The item sets that are both high frequent and high utility can be obtained using the method.
5. The data relationship management is incorporated into the system by tracking the data which is frequent usage of the different kinds of item set.

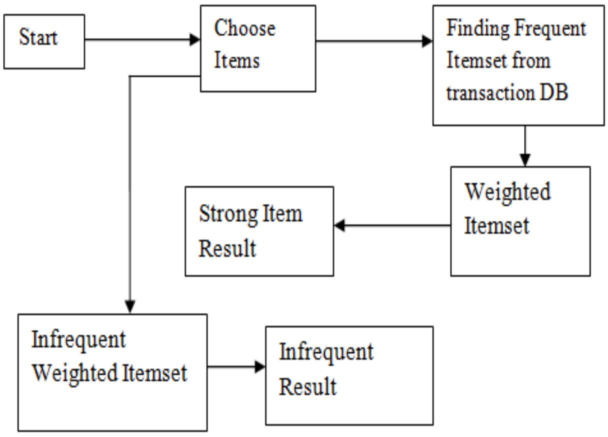
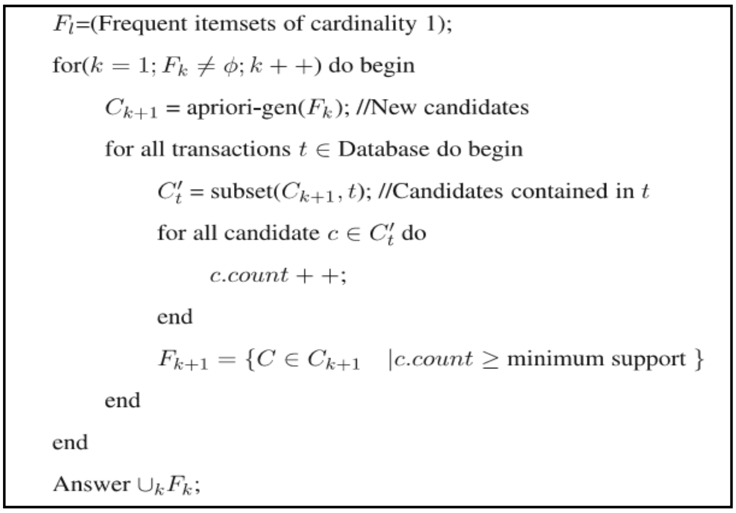

Start
Choose Items
Finding Frequent Itemset from transaction DB
Weighted Itemset
Strong Item Result
Infrequent Weighted Itemset
Infrequent Result

Fig.2 System Flow Model

VII. RANK PRUNING ALGORITHM BASED ON THE APRIORI ALGORITHM

Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level wise complete search algorithm using anti-monotonicity of itemsets, if an itemset is not frequent, any of its superset is never frequent. By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. The impact of the algorithm many of the pattern finding algorithms such as decision tree, classification rules and clustering techniques that are frequently used in data mining have been developed in machine learning research community. Frequent pattern and association rule mining is one of the few exceptions to this tradition. The introduction of this technique boosted data mining research and its impact is tremendous. The algorithm is quite simple and easy to implement. Experimenting with Apriori like algorithm is the first thing that data miners try to do.



```
$F_l$=(Frequent itemsets of cardinality 1);
for($k = 1; F_k \neq \phi; k++$) do begin
    $C_{k+1}$ = apriori-gen($F_k$); //New candidates
    for all transactions $t \in$ Database do begin
        $C'_t$ = subset($C_{k+1}, t$); //Candidates contained in $t$
        for all candidate $c \in C'_t$ do
            $c.count++$;
        end
        $F_{k+1} = \{C \in C_{k+1} \quad | c.count \geq$ minimum support $\}$
    end
end
Answer $\cup_k F_k$;
```

Fig.3 Apriori Algorithm

Let the set of frequent itemsets of size k be F_k and their candidates be C_k. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement.

The most outstanding improvement over Apriori would be a method called FP-growth frequent pattern growth that succeeded in eliminating candidate generation. It adopts a divide and conquers strategy by,
(1) Compressing the database representing frequent items into a structure called FP-tree frequent pattern tree that retains all the essential information and
(2) Dividing the compressed database into a set of conditional databases, each associated with one frequent itemset and mining each one separately.

Pattern growth algorithm works on FP-tree by choosing an item in the order of increasing frequency and extracting frequent itemsets that contain the chosen item by recursively calling itself on the conditional FP-tree. FP-growth is an order of magnitude faster than the original Apriori algorithm.

VIII. RESULTS AND DISCUSSION

The Apriori achieves good performance by reducing the size of candidate sets. However, in situations with very many frequent itemsets, large itemsets, or very low minimum support, it still suffers from the cost of generating a huge number of candidate sets and scanning the database repeatedly to check a large set of candidate itemsets. In fact, it is necessary to generate 2^{100} candidate itemsets to obtain frequent itemsets of size 100.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

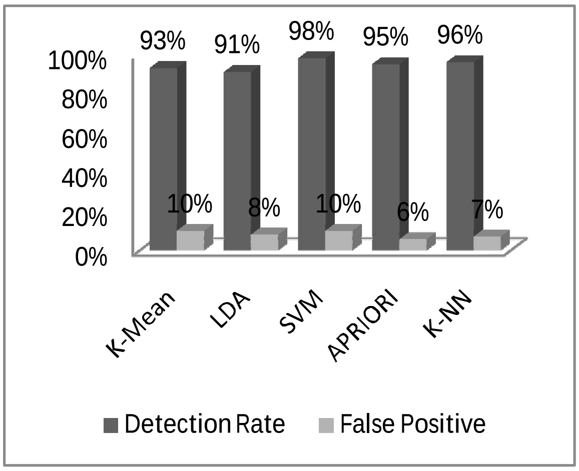


Fig.4 Apriori Comparison with Other Outlier Detection Algorithms

The main aim is to further support the observations that reverse neighbor relations can be effectively applied to outlier detection and Apriori algorithm in both high and low dimensional settings.

IX. CONCLUSION AND FUTURE WORK

In this proposed weighted frequency of occurrence of an item set in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. The frequent and infrequent support min measure, which relies on a minimum cost function, and the frequent and infrequent support max measure, which relies on a maximum cost function, that is the occurrence of an item set in a given transaction are weighted by the weight of the most interesting item. Instead of using any Distance based Outlier algorithm for mining frequent and infrequent records any type of ranking algorithm can be used. The similarity can be checked by other measurement apart from Apriori and will check for the improved performance of outlier detection. In future the usefulness of the novel data mining patterns discovery for outliers from a real life context with the help of a domain expert will possible.

REFERENCES

1. Aggarwal C.C and Yu P.S, “Outlier detection for high dimensional data”, *Proc 27th ACM SIGMOD Int Conf on Management of Data*, 2001, pp. 37–46.
2. Aslam J.A, Kanoulas E, Pavlu V, Savev S and Yilmaz E, “Document selection methodologies for efficient and effective learning-to-rank”, *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2009, pp. 468–475.
3. Breunig M.M, Kriegel H.P, Ng R.T and Sander J, “LOF: Identifying density-based local outliers”, *SIGMOD Rec*, vol. 29, no. 2, pp. 93–104, 2000.
4. Chu W and Ghahramani Z, “Extensions of Gaussian processes for ranking: Semi-supervised and active learning”, *Proc. Nips Workshop Learn. Rank*, 2005, pp. 33–38.
5. Cohn D.A, Ghahramani Z and Jordan M.I, “Active learning with statistical models”, *Proc. Adv. Neural Inf. Process. Syst.*, 1995, vol. 7, pp. 705–712.
6. Doug Wielenga, “Identifying and Overcoming Common Data Mining Mistakes”, *SAS Global Forum Paper* 073-2007.
7. Hautamaki V, Karkkainen I and Franti P, “Outlier detection using k-nearest neighbour graph”, *Proc 17th Int Conf on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 430–433.
8. Lin J, Etter D and DeBarr D, “Exact and approximate reverse nearest neighbor search for multimedia data”, *Proc 8th SIAM Int Conf on Data Mining (SDM)*, 2008, pp. 656–667.
9. Milos Radovanovic, Alexandros Nanopoulos and Mirjana Ivanovic, “Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection”, *IEEETransactions On Knowledge And Data Engineering*, VOL. 27, NO. 5, MAY 2015.
10. Nanopoulos A, Theodoridis Y, and Manolopoulos Y, “C2P: Clustering based on closest pairs”, *Proc 27th Int Conf on Very Large Data Bases (VLDB)*, 2001, pp. 331–340.
11. Papadimitriou S, Kitagawa H, Gibbons P and Faloutsos C, “LOCI: Fast outlier detection using the local correlation integral”, *Proc19th IEEE Int Conf on Data Engineering (ICDE)*, 2003, pp. 315–326.

12. Radovanovi´c M, Nanopoulos A and Ivanovi´c M, “Hubs in space: Popular nearest neighbors in high-dimensional data”, *J Mach Learn Res*, vol. 11, pp. 2487–2531, 2010.
13. Rahm E and Do H. H, “Data Cleaning: Problems and Current Approaches”, *IEEE Bulletin of the Technical Committee on Data Engineering, Vol.23, No.4*.
14. Wang R, Storey V and Firth C, “A framework for analysis of data quality research”, *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995.
15. Zhang K, Hutter M and Jin H, “A new local distance-based outlier detection approach for scattered real-world data”, *Proc 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD)*, 2009, pp. 813–822.
16. Zimek A, Schubert E and Kriegel H.P, “A survey on un-supervised outlier detection in high-dimensional numerical data”, *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

BIOGRAPHY

Shanthini B is a Research Scholar the Computer Science Department, College of Swami Vivekananda Arts and Science, Thiruvalluvar University. She received Master of Computer Application (MCA) degree in 2015 from TACW, Villupuram,Tamil Nadu, India. Her research interests are Data Mining, Frequent Pattern mining, Apriori algorithm, Value Predictions and etc.,