An Outlier Detection Based on Frequent Pattern

Chhanda Ray
State Council of Educational Research & Training
Govt. of West Bengal
Kolkata, INDIA

Pintu Das
Jadavpur University
Kolkata, INDIA

Abstract

Outlier detection and analysis is an important data mining task and in recent years it has been widely used in many practices such as fraud detection, marketing analysis, medical analysis, network intrusion and so on. An efficient outlier detection method on transactional datasets, namely, EFPOR, has been focused in this work. In this algorithm, outliers are detected based on frequent patterns of itemset within transactions. First, the EFPOR algorithm for outlier detection has been described in this work. The time complexity of the EFPOR algorithm has been calculated. At the end of the work, the performance efficiency of the algorithm has been verified by comparing with other algorithms and the experiment results are presented.

• Introduction

Data mining is the process of extracting interesting patterns from large amount of data and it is commonly used in a wide range of profiling practices such as marketing, surveillance, fraud detection and scientific discovery. Outlier detection and analysis is an important data mining task which is used to detect rare and unusual events, deviant objects and exceptions. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. In the field of data mining, there are various useful methods for outlier detection which is usually consists of two different steps. In the first step, the outliers of a given data set are defined and in the next step a suitable method has been find out or developed to detect outliers.

In recent years, outlier detection has drawn significant attention [1, 2, 5, 6, 9, 10, 11, 12, 13, 14] since it has been widely used in many practices such as fraud detection, marketing analysis, medical analysis, network intrusion and so on. Paper [1] proposed Local Outlier Factor (LOF) for each object in the data set indicating its degree of outlier in order to identify density-based local outliers. Since the LOF value of an object is obtained by comparing its density with its neighborhood objects, this method [1] has stronger modeling capability than distance-based outlier detection methods. A novel minimum spanning tree based algorithm for detecting outlier-communities from complex networks is focused in [2]. A new community validation criterion based on the geometric property of data partition of the data set is used in this algorithm [2] in order to find the exact number of communities. In [5], an outlier detection algorithm using Indegree Number, ODIN, has been presented that utilizes k-nearest neighbor (KNN) graph. A clustering-based outlier detection method, namely, CBOD, based on outlier factor of cluster has been illustrated in [6]. In this paper [6], the outlier factor of cluster is defined as weighted sum of distances between clusters. Paper [9] focuses on an outlier detection method for transaction databases using association rules. Another clustering based approach has been introduced in [10] for outlier detection.

In the context of outlier identification, it is generally accepted that if a data object contains more frequent patterns the corresponding data object is unlikely to be an outlier because it possesses the common

features of the dataset. The infrequent patterns that are contained in very few data objects can be descriptions of outliers for a given dataset. A frequent pattern based outlier detection method is proposed in [11]. Paper [12] illustrated an outlier detection algorithm based on clustering analysis. In this paper [12], clustering algorithm is applied initially to obtain the clustering result set and noise clusters are eliminated from the result set based on data count values which are less than certain predefined threshold value. Finally, the outliers are identified based on standard deviation which describes the fluctuation of offset of a given object in a cluster. A new local distance-based outlier detection approach for scattered real-world data is introduced in [13] while [14] focuses a method for outlier detection based on clustering approaches.

Many researchers [3, 7, 8] have also been devoted for outlier removal in order to make the given datasets noise free. In [3], an outlier removal algorithm has been proposed which is a merging of hierarchical, partition, and density-based methods. This algorithm [3] uses a sparse graph in which nodes represent data items and weighted edges represent the distances between the data points. A new algorithm, namely, Outlier Removal Clustering through Minimum Spanning Tree (ORCMST), is introduced in [7] which do not require a predefined cluster number. This algorithm [7] works in two different phases. The first phase focuses on the construction of the minimum spanning tree and the second phase deals with the outlier detection and removal. In [8], an distance-based outlier removal method has been represented which is a combination of three partition-based schemes such as Partitioning Around Medoid (PAM), Clustering Around Large Applications (CLARA), Clustering Large Applications Based on Randomized Search (CLARANS), and k-mediod clustering algorithm. Paper [4] introduces an approach for searching frequent patterns without candidate generation.

Most of the outlier detection methods discussed above is applicable for numerical data only. Moreover, in the context of discrete and transactional data, the performance of these algorithms is not good. Further, only a very limited number of studies have attempted to detect outliers from categorical data such as record data having nominal attribute values and transactional data. In this paper, an efficient outlier detection method based on frequent patterns of itemset within transaction, EFPOR, has been introduced for transactional data. An outlier detection method based on frequent pattern has also been proposed in [11], but the major drawback of this algorithm is that when transactions outlier measure (FPOF value) is calculated by using this approach, the support degrees of the corresponding transactions itemset and all the subsets of the itemset are both added. This duplicate computing would enlarge the differentials of normal degrees among transactions. In this approach [11], the outlier measure *FPOF* cannot reflect the normal degree of transactions accurately. In our work, the outlier measure has been redefined and the duplicate added problem has also been solved as well in order to detect outlier more precisely. Thus, our work differs from [11] in a major way.

The organization of the paper is as follows. The algorithm for outlier detection of transactional datasets based on frequent pattern is illustrated in Section 2. In Section 3, the complexity of the outlier detection algorithm, EFPOR, has been presented. Section 4 focuses on the detailed experimental results and the work is concluded in Section 5.

• Algorithm for Frequent Pattern Outlier Detection

The concept behind the frequent pattern based outlier detection method is that if a data object contains more frequent patterns, it is unlikely to be an outlier. Let T is a transactional dataset and t is a transaction such that t T. Let also assume that X = {I_1, I_2, I_5}, I_i I, where I is the itemset of T and X is a frequent pattern of frequent pattern set F, X F. Hence, the non-empty subsets of X are {I_1, I_2, I_5}, {I_1, I_2}, {I_1, I_5}, {I_2, I_5}, {I_1}, {I_2}, and {I_5}. Let t contains X, thus, t X, and t also contains all the subsets of X. In this case, since X is a frequent pattern, the subsets of X must be frequent patterns too. It is important to note that in the frequent pattern mining process, a superset frequent pattern that contains more items (has long length) is combined by subset frequent patterns that contain less items (has short length).

Therefore, the longer superset frequent pattern has more subset frequent patterns. A transaction that contains longer superset frequent patterns is more likely to be a normal transaction because it has more subset frequent patterns than other transactions. In contrast, a transaction that contains short frequent patterns is more likely to be an outlier. Based on the above concept, a new measure for outlier detection, LEPOF, which reflects the normal degree of a transaction, is defined as follows.

LFPOF(t) = --------- (i)

In this case, we assume that F is a complete set of frequent itemsets, F(t) is a frequent pattern set which composes of frequent patterns contained in transaction t. Hence, F(t) is derived in the following way.

F(t) = {X|X F t X }

Let X_{max} is the longest frequent pattern in F(t), $|X_{max}|$ represents the length of X_{max} and |t| represents the length of transaction t. The numerator in equation(i) shows that a transaction contains longer frequent pattern is more normal.

In this work, the *Apriori* algorithm has been used first to generate the frequent pattern set F and then for each transaction t in the dataset T, the frequent pattern set F(t) is generated. The longest frequent pattern X_{max} in F(t) has determined thereafter and for each transaction t, normal degree LEPOF(t) is computed. All the transactions are sorted in ascending order based on their EFPOR value and top k frequent pattern outliers are detected at the end. The outlier detection algorithm, EFPOR, is depicted in the following.

Algorithm EFPOR

Input: Transaction dataset T, user-defined threshold minimal support *min_sup*, user-defined threshold value for top *k-fp-outlier*.

Output: top k outlier transactions with the corresponding k lowest LEPOF values.

Begin

1: Generate the frequent pattern set F on database T by using Apriori algorithm with the minimum support value *min_sup*.

2: For each transaction t T

3: LEPOF(t) = 0

4: For each frequent pattern X F

5: if t X

6. add frequent pattern X in the set F(t)

7: end if

8: end for

9: Find the longest frequent pattern X_{max} in F(t)

10: Calculate LEPOF(t) = $|X_{max}|$ / |t|

11: end for

12: Sort the transactions in ascending order based on their LEPOF value

13: Output the top k transactions as outliers using *k-fp-outlier*

14: end

- **Performance Analysis for EFPOR Algorithm**

In this Section, the complexity of frequent pattern outlier detection algorithm has been illustrated. The computational cost of EFPOR algorithm is mainly consists of two parts. The first part includes the cost of generating frequent pattern set by using *APriori* algorithm while the second part includes the cost of executing the inner loops from line no. 2 to 9 in the

EFPOR algorithm. The time complexity of generating frequent pattern set using *Apriori* algorithm increases as the size of the transaction dataset and the width of the itemset within the corresponding transaction dataset increases. The computational time of the *Apriori* algorithm would be quite long when the size of the transaction dataset and the corresponding width of the itemset are extremely large. The nesting of for loops from line

no.2 to line no. 9 in the EFPOR algorithm causes O(|T| × |F|) asymptotic time complexity. When |T| and |F| are getting larger, the execution time of the EFPOR algorithm is getting longer. However, the time complexity O(|T| × |F|) is acceptable as long as |T| and |F| are limited in certain scale.

• Experimental Results

In this Section, the algorithm EFPOR has been executed on several transactional datasets for detecting accuracy of outlier detection method based on frequent pattern. In order to compare the efficiency of EFPOR algorithm, the experiments are made on both the FP algorithm [11] and EFPOR algorithm and outcomes are analyzed. The experimental results of EFPOR algorithm are shown in the following.

Items	Transactions	Minsup	Outlier Factor	Time Taken (seconds)
8	100	0.4	0.3	0.015
8	500	0.4	0.3	0.015
8	1000	0.4	0.3	0.016
8	2000	0.4	0.3	0.031
8	5000	04	0.3	0.047
8	10000	0.4	0.3	0.094
8	15000	0.4	0.3	0.156
8	20000	0.4	0.3	0.25
8	25000	0.4	0.3	0.39
8	30000	0.4	0.3	0.547
8	35000	0.4	0.3	0.735
8	40000	0.4	0.3	0.937
8	45000	0.4	0.3	1.172
8	50000	0.4	0.3	1.438

In the above experiments, the total number of 8 items has been taken in the itemset with different number of transactions. However, the number of items in the itemset can vary. The efficiency of EFPOR algorithm in comparison with FP algorithm [11] is depicted in the following figure.

The above experimental result shows that the EFPOR algorithm is more efficient than the FP algorithm.

• Conclusion

In this paper, an efficient algorithm, EFPOR, has been focused for outlier detection on transactional datasets based on frequent patterns of itemset within transactions. The time complexity of the outlier detection algorithm has been analyzed. It is obvious from the experimental results that the performance of the EFPOR algorithm is more efficient than other existing algorithms. However, the Apriori algorithm is used in this work for frequent pattern generation due to its simplicity. Moreover, transactions with itemset of width 8 have been considered in our experiments.

In future scope, efficient frequent pattern generation algorithm can be used for improved performance. Moreover, the parameter sensitivity problem of EFPOR method can be solved and it can be figure out how to fix the most proper *min_sup* for EFPOR method. Further, in order to increase efficiency, improved data structures can be used in future work. In order to verify the performance accuracy of the EFPOR algorithm, it can be apply in other application areas such as fraud detection, medical analysis, network intrusion, marketing analysis and so on.

• References

[1] M. M. Breunig, H. P. Kriegel, R. Ng, J. Sander, "LOF: Identifying Density-Based Local Outliers," Proceedings of ACM SIGMOD Conference, 2000, pp150-165.

[2] S. Chidambaranathan, S. John Peter, "Detection of Outlier-Communities using Minimum Spanning Tree", Proceeding of the Journal of Emerging Trends in Computing and Information Sciences, October 2011, Vol.2, No. 11.

[3] A.M. Fahim, G. Saake, A. M. Salem, F. A. Torkey, M. A. Ramadan, "DCBOR: A Density Clustering Based on Outlier Removal", Proceedings of World Academy of Science, Engineering and Technology 45 2008.

[4] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proceedings of Data Mining and Knowledge Discovery, 2004, pp. 1-8.

[5] V. Hautamaki, I. Karkkainen, P. Franti, "Outlier Detection Using k-Nearest Neighbor Graph", Proceedings

of the International Conference on Pattern Recognition, Cambridge, UK, August 2004, Vol.3, pp430–433.

[6] S. Jiang, Q. An, "Clustering-Based Outlier Detection Method", Proceedings of 5th International Conference on Fuzzy Systems and Knowledge Discovery.

[7] T. Karthikeyan, S. John Peter, "Outlier Removal Clustering through Minimum Spanning Tree", Proceedings of International Journal of Computer Applications, October 2011, Vol. 31 No.10.

[8] P Murugavel, M. Punithavalli, "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", Proceedings of International Journal on Computer Science and Engineering (IJCSE), January 2011, Vol. 3, No. 1.

[9] K. Narita, H. Kitagawa, "Outlier Detection for Transaction Databases using Association Rules," Lecture Notes in Computer Science, 2008, pp1-7.

[10] R. Pamula, J. K. Deka, S. Nandi, "An Outlier Detection Method Based on Clustering", Proceedings of 2nd International Conference on Emerging Applications of Information Technology, 2011.

[11] H. Zengyou, X. Xiaofe, "FP-Outlier: Frequent Pattern Based Outlier Detection," Proceedings of Computer Science and Information Systems, 2005, pp. 1-6.

[12] Y. Zhang, J. Liu, B. Song, "A New Algorithm for Outlier Detection based on Offset", Proceedings of 5th International Conference on Information Assurance and Security, 2009.

[13] K. Zhang, M. Hutter, H. Jin, "A new local distance-based outlier detection approach for scattered real-world data", Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2009, pp813–822.

[14] Moh'd Belal Al Zoubi, "An Effective Clustering-Based Approach for Outlier Detection", Proceedings of European Journal of Scientific Research, Vol. 28, No.2, 2009, pp310-316.