ARTICLE OPEN ACCESS

# AN EFFICIENT ALGORITHM FOR DETECTING OUTLIERS IN A DISTRIBUTED ENVIRONMENT USING MINIMAL IN-FREQUENT ITEM SET PATTERN MINING

**Chandra Ravi Chandran[1] and Ajitha Padmanabhan[2]**

[1]*Dept. Of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, INDIA*
[2]*Dept.of M.Sc [SS&CS], K. G College of Arts and Science, Coimbatore, Tamil Nadu, INDIA*

## ABSTRACT

*Outlier's detection in a distributed data surges the classification and prediction of the tasks easier and accurate. Outliers are unusual patterns that occurs rarely and has less incidence in the data Distance based and density based algorithms exists in literature. The In- Frequent item set mining can be utilized to detect outliers which may increase the performance in terms of accuracy. An efficient algorithm is proposed in this paper to detect outliers by defining a certain minimum support threshold for identifying outliers in distributed data by mining minimal in-frequent patterns in the data.*

**Published on:** 08th– August-2016

KEY WORDS

*Outliers; In-Frequent Pattern; distributed data mining;*

***Corresponding author: Email:** ajitha.mca@gmail.com; **Tel.:** +91-9843862331

## INTRODUCTION

Outliers are an unusual pattern that occurs rarely and has fewer incidences in the data and normally have lesser support [1]. When the data are from various sources and distributed, there are chances for existence of outliers. Detecting of outliers in the disseminated environment is highly challenging and has very few explorations in this arena [6]. Outliers are dissimilar or inconsistent data that deviates from the normal with smallest measurement [8]. Distributed Data Mining is mining data from different and various sources. Perceiving outliers from datasets has many applications like credit card fraud detection, medical diagnosis, market segmentation and e-commerce[4].

## RELATED WORK

Existing methodologies exist for detecting outliers in a centralised environment frequent pattern mining is used for outlier's detection a distributed data without candidate generation [3]. There are different type of approaches for outliers detection like distance based, density based, clustering based and distribution based. Artificial Intelligence Based approaches to outlier detection like Support Vector methods, fuzzy logic based methods, Genetic algorithm based methods are also available in the literature [14].

Frequent pattern item set detects outliers and assign outlier score to each data point based on the frequent item set it contains. Most of the existing literature shows only frequent item set mining, which may be easier to eliminate outliers. Basically, discovering infrequent patterns in the data sets are considered as outliers. Outliers itself is the attributes that are minimally consistent with the pattern of the data [9].

Mining in-frequent items is proposed in a algorithm called AfRIM [11]. The in-frequent items are searched in top-down manner but with minimum or zero support. MSApriori algorithm is proposed [7] to identify the in-frequent item set based on the high confidence rules and multiple support thresholds which decreases the efficiency. Multiple support thresholds considers data sets of individual nature and to be provided for each and every data sets separately which may in-turn reduces the efficiency [9,10].

Mining frequent item sets are identified in the association rules for fast discovering the frequent item sets so that occurrence of data are considered[12] other in-frequent are not considered and discarded as outliers.Confabulation –inspired Association Rule Mining (CARM)[3] discussed mining both frequent and infrequent pattern set mining inspired on cogency based approach. In-frequent item set discovery by single pass through the association rule datasets.But this approach is based on the conditional probability that exists.Outlier detection using distance based, density based, frequent patterns, density based, distance based , Artificial neural networks, information theoretic based approaches[13][5].

## METHODOLOGY

Integrating In-frequent pattern mining for outlier detection is of a novel approach as it interestingly offers high accuracy of outlier's discovery in vast amounts of data. This paper discusses the outlier detection in distributed sources. Current literature shows detecting outliers in distance based and density based outlier's detection. A new methodology is discussed here to detect outliers. In-Frequent item set pattern discovery in outliers by having automatically assigning a parameter to the mini-support. Secondly finding closed frequent item sets to reduce the memory if the data sets are of large nature with a minimum support threshold.

**Algorithm : CiFPMDiscover**

```
1.Input the data from various sources
2.Identify all frequent itemsets and generate individual candidates that are not discovered.
3.Frequent Pattern Support is calculated to check whether superset for the same support as frequent patterns exists or not.
4.If FPSupp =  SuperSupp then
        $iFP=D=\{i_1,i_2,\ldots i_n\}$
       else
       iFP={NULL}
5.Till all the iFP=NULL
6.Iterate till all the possibilities of  super set checked with other MinSupp
7.CiFPM is generated when no supersets of same support count
8.Terminate all the item set generation
9.MinSupp={α}
10.If MinSupp then OutDet
 11. Terminate the process
```

The algorithm Closed in-Frequent Pattern Mining Discover is used to discover outliers and discard it when there is no superset that has same support count as the original itemset. It increases accuracy in finding outliers with single pass so outliers can be easily found.

## RESULTS

CiFPMDiscover algorithm , finds in-frequent pattern mining with closed itemsets so that it provides minimal space to find outliers. The datasets considered are BreastCancer Winscoin datasets.

**Table: 1.Class Distribution of Wisconsin Cancer Breast Cancer Dataset**

| Case | Class codes | Number of instances |
|---|---|---|
| Commonly occurring classes | 2 | 65.5% |
| Rare class | 4 | 34.5 |

**Table-1** shows the class distribution of Wisconsin breast cancer datasets. Commonly occurring classes shows the normal classes and rare class shows the outliers in the datasets.When comparing the CiFPMDiscover with other algorithms of FPOF,CBLOF the detecting of outliers is shown below.

**Table- 2** shows the minimum support threshold for identifying outliers in having minimum support threshold for breast cancer datasets in benchmarked UCI machine repository datasets. If the minimum threshold of α is reached the dataset is considered as outliers and they are discarded.

**Table: 2. comparision of proposed CiFPMDiscover to FPOF, CBLOF**

| Number of Records | Number of Outliers Detected | | |
|---|---|---|---|
| | FPOF | CBLOF | CiFPMDiscover |
| O | 0 | 0 | 0 |
| 4 | 3 | 4 | 3 |
| 8 | 7 | 7 | 6 |
| 16 | 14 | 14 | 11 |
| 24 | 21 | 21 | 18 |
| 40 | 31 | 32 | 30 |
| 48 | 35 | 35 | 35 |
| 56 | 39 | 38 | 36 |
| 64 | 39 | 39 | 36 |
| 72 | 39 | 39 | 38 |
| 80 | 39 | 39 | 38 |
| 100 | 39 | 39 | 38 |
| 112 | 39 | 39 | 39 |

**Table: 3 .Execution times in respect to the centralized algorithm**

| Dataset/l | 5 | 10 | 15 |
|---|---|---|---|
| Breast Cancer | 230.1 | 126.4 | 96.5 |
| Poker | 210.1 | 112.3 | 83.3 |
| Cov Type | 230.1 | 126.4 | 96.5 |

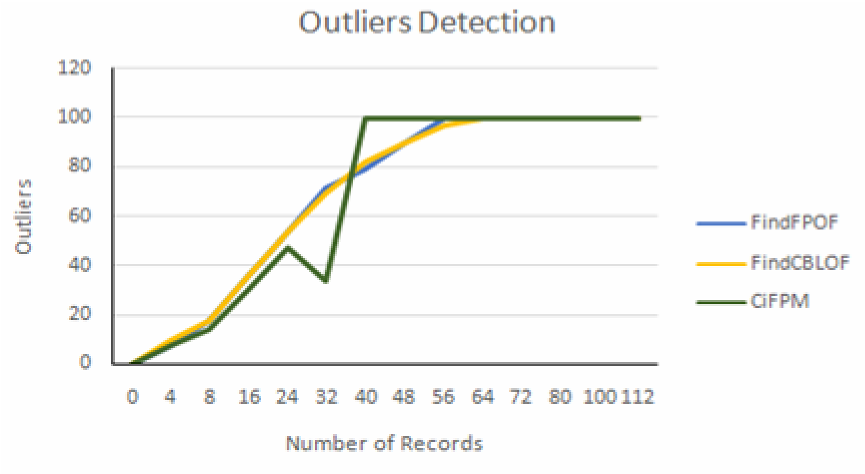

**Fig: 1.Comparision of Number of Outliers detected**

……………………………………………………………………………………………………..

The **Figure- 1**, shows the comparision of number of outliers detected from the proposed cifpm to the other algorithms for the datasets, breast cancer, poker and ecotype data sets available in uci machine repository.

## CONCLUSION AND FUTURE WORKS

CiFPMDiscover algorithm detects outliers with using minimum support threshold. Using the closed in-frequent pattern detection by discarding the attributes that does not support with minimum threshold limit. The proposed algorithm deals with single pass in datasets and saves in memory limitage. Accuracy and memory requirements that considered for discovering outliers is comparatively efficient then the existing methods. Detection of outliers in distributed data sources can be further extended to domain based outlier detection. Automatic detection of outliers based on the dataset may be also explored further.

## CONFLICT OF INTEREST

None declared.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

No financial support was received for this work.

## REFERENCES

[1] Adda M, Wu L, Feng Y. [2007] Rare itemset mining. In Proceedings of the 6th International conference on machine learning and applications (ICMLA ‘07) (pp. 73–80). Washington, *DC: IEEE Computer Society*.

[2] Agrawal.R, Srikant.[1994] Fast Algorithms for Mining Association Rules. In Proceedings of VLDB’94, 478-499,

[3] AzadehSoltaniandM.-R.Akbarzadeh-T.[2014]onfabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets, *IEEE* transactions on neural networks and learning systems, Cateni

[4] SV Colla,[2013] Data Processing for Outliers Detection, Pattern Recognition: Methods and Application, iConcept Press Ltd, ISBN: 978-1-922227-08-9.:1-21.

[5] Chandola V, Banerjee A, Kumar V. [2009]. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3):15.

[6] Chandra.E, P Ajitha,[2015] An Algorithm For Detecting Outliers In Distributed Environment” International Journal of Applied Engineering Research ISSN 0973-4562 10(1): 1519-1523 © *Research India Publications.*

[7] Han JJ, Pei Y Yin and R Mao.[2004] Mining frequent patterns without candidate generation: a frequent-pattern tree approach, *Data Mining and Knowledge Discovery*, 8(1): 53–87.

[8] Hawkins.D.[ 1980] Identification of Outliers. Chapman and Hall, Reading, London, ISBN 978-0412219009

[9] He Z, Xu X, Huang, JZ, Deng S.[2005] FP-Outlier: Frequent Pattern Based Outlier Detection. *Computer Science and Information Systems*, 2( 1): 103-118.

[10] Jure Leskovec,Anand Rajaraman, Jeffrey D.Ullman. [2014] Mining of Massive Datasets. Cambridge University Press. ISBN 978-1107015357.

[11] Liu B, Hsu W, Ma Y. [1999] Mining association rules with multiple minimum supports. In Proceedings of 5th ACM SIGKDD international conference on knowledge discovery and data mining (KDD ‘99) (pp 337–341). New York

[12] Moens.S, Aksehirli E, Goethalsn.B. Frequent Itemset Mining for Big Data”, *IEEE International Conference* on Big Data,.111-118.

[13] Pimentel MA, Clifton DA, Clifton L, Tarassenko LA.[ 2014] review of novelty detection. Signal Processing, 99:215–249.

[14] Rasheed, F, Alhajj, R. [2014] A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences,*IEEE Transactions on Cybernetics*, 44( 4): 569-582.

## ABOUT AUTHORS

*Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 15 yrs of experience in teaching including 6 months in the industry. Presently she is working as Professor, Department of Computer Science in Bharathiar University, Coimbatore. She has published more than 50 research papers in National, International Journals and Conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. She is Life member of CSI and editor in various International Journals.*

*P.Ajitha,, received her B.Com from Bharathiar University, Coimbatore in 1998 and received MCA from Bharathiadsan University, Trichy in 2001. She obtained her M.Phil in the area of Data Mining in 2004. She has 9 years of experience in teaching and 3 months of industrial experience. Currently, she is working as Assistant Professor, Department Of M.Sc Software Systems and Computer Science, K.G College of arts and science, Coimbatore and pursuing her Ph.D in Bharathiar University,Coimbatore. She has presented more than 10 research papers in National and International Conferences and published mre than 5 papers in an various International Journals. Her Research Interest lies in Distributed Data Mining, Machine Learning and Artificial Intelligence. She is Life a member of CSI, life member of Institute of Advanced Scientific Research and also a member IASCIT.*