Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

An Integration Approach to Detect Outliers in A Distributed Environment Using A Novel Approach

E.Chandra[1] and P.Ajitha[2*]

[1]Department of Computer Science, Bharathiar University, Coimbatore
[*]Department of Software Systems & Computer Science(PG), KG College of arts and science, Coimbatore, Tamil Nadu, India ajitha.mca@gmail.com,9843862331

ABSTRACT

Outliers are the observation that deviates from the norm. Detecting outliers is a challenging task. When the datasets are massive and scalable the preprocessing tasks are essential for higher accuracy in prediction. This paper paves way to preprocess the data through identifying outliers through various algorithms Classify Purgeout is the proposed algorithm that integrates the data from various sources and detect outliers. Existing algorithms detects outliers in an centralized environment, which does not consider preprocessing aspects. The algorithm proposed here, eliminates the necessity of data processing in an centralized environment, which provides efficient classification and prediction accuracy.

[*]Corresponding author

P. Ajitha

Assistant Professor,

Department of Software Systems & Computer Science(PG),

KG College of arts and science, Coimbatore

Email: ajitha.mca@gmail.com,9843862331

1. INTRODUCTION

Existing methodologies for detecting outliers in a centralised environment are frequent pattern mining that is used for outlier's detection in 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.

Frequent pattern [4] 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.

Mining in-frequent items is proposed in AfRIM[5]. The in-frequent items are searched in top-down manner but with minimum or zero support. MS A priori algorithm is proposed[6] 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.

Mining frequent item sets are identified in the association rules for fast discovering the frequent item sets so that occurrence of data are considered[7] other in-frequent item sets are not considered and discarded as outliers. Confabulation–inspired Association Rule Mining (CARM)[8] discussed both frequent and infrequent pattern set 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 based, information theoretic based approaches are discussed in[9].

2. REVIEW LITERATURE

This paper, offers the heuristics that can be applied to the classifying outliers , which was not considered in the existing algorithms. Existing Outlier detection techniques that were discussed are not suitable for high dimensional data sets. Evolutionary algorithms[10] can also be one of the possibilities for detecting outliers in very high dimensional datasets for the disseminated environment.

Most of the Outlier detection approaches are statistical based, deviation based, density based or distance based. Very few of the existing outlier detection techniques uses the distributed data mining techniques like Distributed Clustering, Distributed Association Rules Mining and Distributed Decision Trees and Support Vector Machines. Distributed Data Mining contains large high dimensional data containing outliers, which may affect the overall classification, prediction tasks. Of course, parallel data mining is also possible but due to some inherent disadvantages like requiring maximum resources, data mining strategies are used for it. Especially when dealing with detecting of outliers for large streams of data, distributed data mining comes in handy with its techniques. Distance based outliers are also considered for streams of data [11].

2.1 One-Class Classification – Unsupervised Approach

Unsupervised or unlabeled learning approaches for network anomaly detection have been recently proposed [12]. These methods aim to work on datasets of traffic extracted from real networks without the necessity of a labeling process. Unlabeled anomaly detection systems[13] are based on the reasonable assumption that the percentage of attack patterns in the extracted traffic traces is usually much lower than the percentage of normal patterns. Furthermore, it is possible to use signature-based IDS in order to filter the extracted traffic by removing the known attacks, thus further reducing the number of attack patterns possibly present in the dataset.

Another assumption is that the attack patterns are supposed to be distinguishable from the normal patterns in a suitable feature space. The term “unlabeled anomaly detection” used in the intrusion detection field actually refers to what in machine learning is more often called “novelty detection”, “outlier detection” or “one-class classification”. One-class classification algorithms pursue concept learning in absence of counter examples. The objective of one-class classification is to find a decision surface around the target objects, (i.e., the normal traffic, in case of network anomaly detection) so that patterns that lie inside this decision surface are classified as targets (i.e., normal traffic), whereas patterns that lie outside are classified as outliers (i.e., anomalous traffic).

2.2 Support Vector Machines

Support Vector Machines (SVM) [14] are one of the most actively developed classification and regression methodologies with its margin maximization and systematic non-classification via kernel tricks. Because of the issues like scalability, applicability and interoperability the SVM are used less. The issues of SVM and approach to handle the challenges of SVM in outlier detection in respect to the high dimensional data is eliminated in the proposed research work.

Training SVMs on large datasets is still a challenging problem. Sample reduction methods have been proposed and shown to reduce the training complexity significantly, but more or less trade off the generalization performance. An efficient sample reduction method for multi-class classification using one-vs-rest SVMs, called MUlti-class Sample Selection (MUSS). For each binary one-vs-rest classification problem, positive samples and negative samples are selected based on the distances from the cluster centres of positive class, assuming that positive samples with large distances from the positive centres and negative samples with small distances from the positive centers are near the classification boundary. The intention of clustering is to improve the computation efficiency of sample selection, other than to select from cluster centres as previous methods MUSS, to select boundary samples as training data to substantially reduce the scale of a multi-class dataset and effectively speed up the training of SVM models. With the one versus rest (o-v-r) style of multi-class SVMs, MUSS gets a set of selected samples for each o-v-r SVM model. For a certain class cas positive class, MUSS gets a predetermined number of cluster centres of class c by some clustering algorithm. And then, with these centres as reference points MUSS can sharply and efficiently reduce the scale of the training set. Meanwhile, the possible serious imbalance of training data caused by o-v-r style can be controlled [15].

SVM in combined with other machine learning techniques yields better results and performance. Fast nearest neighbour condensation with SVM is discussed in the literature which uses two techniques decomposition and data reduction [10]. Clustering Based SVM are also discussed to embrace the issues of scalability and interoperability. CB-SVM presents better performance and accuracy [156]. For gene selection in cancer classification the SVM are able to use with space dimensionality and feature selection. SVM recursive feature elimination is also discussed in the literature. The distributed method for detecting distance-based outliers in very large data sets was proposed [12]. Detecting outlier by solving set, which is a small subset of the data set that can be also employed for predicting novel outliers. The method exploits parallel computation in order to obtain vast time savings. Indeed, beyond preserving the correctness of the result, the proposed schema exhibits excellent performances. From the theoretical point of view, for common settings, the temporal cost of our algorithm is expected to be at least three order of magnitude faster than the classical nested loop like approach to detect outliers.

2.3 Clustering Based SVM

To train a very large data set Clustering Based-SVM(CB-SVM) can be used. Basic idea of this is using the hierarchical micro-clusters (i.e., CF tree). Micro-clusters trees are constructed with

one-scan of data with the limited resources available. CF tree is called as Clustering Feature (CF) is utilised to build tree with minimal scans in dataset. To construct an accurate Support Vector Machine(SVM) boundary function the CF tree are used. Selective sampling otherwise called as active learning can be used. Selecting the data which maximizes the benefit of learning is basic idea of selective sampling. In the SVMs, the selective sampling makes use of the data which are close to the boundry in the feature space. After the selection, the accumulation of the low margin data nearest to the Support Vectors(SV) of boundry are considered to the next round. In each round that are close to the boundary in the feature space, the data have higher chances to become the SVs of the boundary for the next round [16].

Finer and coarser samples near and farther the boundaries are considered. Splitting the group as near and far enhances the best data to be considered. In this way, the SVs are induced. Decluster of SVs helps to keep the total number of training data points as small as possible. Active learning, scans only the needed data randomly, by CF tree with attributed based boundry.

While selective sampling needs to scan the entire data set at each round to select the closest data point, CB-SVM runs based on the CF tree which can be constructed in a single scan of the entire data set and is carrying the statistical summaries that facilitates constructing an SVM boundary efficiently and effectively. Resources needed for the CB-SVM is considerably reduced as one-scan is utilised.

The sketch of the CB-SVM algorithm follows:

1. Positive and Negative data set are used for creating two Clustering Feature Tree.
2. These two CF trees are constructed independently.
3.SVM boundary function from the centroids of the root entries are trained. The trained node contains only the few entries in the root node – constructed two CF trees. If the root node contains too few entries , the entries of the nodes in the second levels of the trees are considered for training.
4. The next level entries near the boundry are declustered.
5. Children entries are declustered from the parent entries. These declustered entries are accumulated into the training set with the non-declustered parent entries.
6.Another SVM is constructed from the centroids of the entries in the training set, and repeated from step 3 until nothing is accumulated.
7. Till the entire dataset is constructed, the trees are created.

Figure 1: CB-SVM Algorithm

The CF tree is a suitable base structure for CB-SVM to perform the selective declustering efficiently. The clustered data also provides better summaries for SVMs than random samples of the entire data set because the random sampling is susceptible to a biased (or skewed) input, and thus it may generate undesirable outputs especially when the probability distributions of training and testing data are not similar, which is common in practice.

The Clustering Based-SVM is utilised to generate outliers when inputs are of skewed in nature. High data sets of dimensional nature can also be clustered and declustered efficiently based on the type of datasets.

3. PROPOSED METHODOLOGY\

The algorithm Closed In-Frequent Pattern Mining Discover (CiFPM) [1] discovers outliers and discard it when there is no superset that has same support count as the original item set. It increases accuracy in finding outliers with single pass so outliers can be easily found.

- Algorithm Classify Purge-Out
- Min Entropy Calculated based on One-Class heuristic
- Classify the data
- Min Supp calculated in Ci FPM Discover
- Purge outliers based on calculation.
- Iterate till all outliers purges
- Terminate the Process.

Figure 2: Algorithm Classify-Purge out

On finding the minimum entropy based threshold the outliers are sensed and mined from the various nodes. After the entropy based classification of the data sets that are determined by heuristic search and discovering in-frequent based mining with closed item set generation the outliers are detected. On generating outliers in the data sets by One-Class Heuristic[2] and CiFPM Discover, a model is build where the instances in the data sets are classified as outliers or normal. Classify-Purge Out Model involved for classifying the data sets that are normal and outliers by mining in large data sets and outliers are purged. Classifying from the local sites in nodes and generating the instances as normal and outliers.

The One-Class Heuristic Outlier algorithm, takes input as datasets and finds outliers (supposed to be *k*) and select points as outliers with thorough heuristic manner. Initially, set of outliers is specified as empty and all points are marked as non-outlier. Then, scan the datasets and

select k points as outliers. This algorithm takes data sets and applies heuristic search for the entire datasets.

One-Class Heuristic Outlier algorithm , considers datasets and t as record and each record is labeled as non-outlier and hash tables for attributes is also constructed and updated. First scan the datasets for k times to find exact k outliers .i.e. one outlier is identified and defined in each pass. In each scan , each record t is labeled as nonoutlier and later changed to outlier , entropy is computed for before and after the outlier identification , changed entropy computation is re-calculated. Records that achieves maximal-entropy impact is selected as outlier in each scan and added to set of outliers. Use of hashing technique, *O(1)* frequency can be calculated with attribute values. Decreased entropy value *O(n)* expected time, since changed value is dependent on attribute values of the record to be temporally removed. The One-Class Heuristic Outliers is compared with LDSS and DSS for their accuracy and speed up. The datasets considered are of G3d,Covtype,Poker,2Mass.

4. RESULTS AND DISCUSSION

CiFPM Discover algorithm, finds in-frequent pattern mining with closed item sets so that it provides minimal space to find outliers. The datasets considered are Breast Cancer Wins coin datasets.

Table :1 Class Distribution of Wisconsin Cancer Breast Cancer Dataset

Case	Class Code	No. of Instances
Commonly Occuring classes	2	65.5%
Rare Class	4	34.5%

The above table 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 with other algorithms of MIFP, FP the detecting of outliers is shown below.

Table 2: Detecting No. of outliers—sample

Top Ratio (No. of Recs)	No. of Outliers detected		
	FPOF	CBLOF	CiFPM
0%(0)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1%(4)	3 (7.69%)	4 (10.26%)	3 (7.69%)
2%(8)	7 (17.95%)	7 (17.95%)	6 (15.38%)

Table 6.3: Execution times in respect to the centralized algorithm

Dataset/l	5	10	15
Breast Cancer	230.1	126.4	96.4
Poker	210.1	112.3	83.3
Covtype	230.1	126.4	96.5

This table, shows the comparison of the combination of all the algorithms and with the proposed algorithm of the classify purge out with the centralised algorithm of sequential decision tree which discussed earlier in this thesis. Comparison and analysis with the generalised central algorithm along with the benchmarked data sets are used.

Table 6.3 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. Based on the One Class Heuristic to calculate the Entropy with

$$E(X) = \sum_{i \in D(x_i)}^{n} O_i$$

$$E(Y) - \frac{E(x)}{n}, E(Y) - \alpha$$

for considering or calculating the outliers that exists in the data from the local sites. The Entropy of the E(X) is calculated and based on the entropy calculation heuristic is determined in local sites. On Heuristic search, the outliers are determined which does not meet the determined value. Determination of the heuristic search with one-class values is defined for the each search on the data sets.

CiFPM Discover discovers outliers based on the in-frequent item sets that are closed. Generation of in-frequent pattern are mined in the large volumes of data. Closed in-frequent items sets are generated and minimum support are calculated. If the item set generated has greater or equal to the minimum support then it is considered as closed and no super sets exists for that item sets generation that is considered for further generation of candidate items sets. Support count is calculated and threshold for the minimum support or support count is referenced with the item sets generation. The closed frequent item sets are discarded if it reaches minimum support count.

On finding the minimum entropy based threshold the outliers are sensed and mined from the various nodes. After the entropy based classification of the data sets that are determined by heuristic search and discovering in-frequent based mining with closed item set generation the outliers are detected. On generating outliers in the data sets by One-Class Heuristic and CiFPM Discover, a model is build where the instances in the data sets are classified as outliers or normal. Classify-Purge Out Model involved for classifying the data sets that are normal and outliers by mining in large data sets and outliers are purged. Classifying from the local sites in nodes and generating the instances as normal and outliers.

5. CONCLUSIONS AND FUTURE WORKS

CiFPM Discover 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 limit. In Classify-Purge Out, the outliers are purged so that accuracy and memory required for purging outliers is comparatively efficient than 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 domain may be also explored further. Comparison and analysis with general centralised algorithm along with the benchmarked datasets are discussed in this paper

REFERENCES

1. Chandra Ravi Chandran1 and Ajitha Padmanabhan, "An Efficient Algorithm For Detecting Outliers In A Distributed Environment Using Minimal In-Frequent Item Set Pattern Mining", SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS),*ISSN: 0976-3104,The IIOABJ, 2016.*
2. Dr. E. Chandra, P. Ajitha, "An Algorithm For Detecting Outliers In Distributed Environment", International Journal of Applied Engineering Research, ISSN 0973-4562.Research India Publications2015;10(1): 1519-1523
3. He. Z., Xu. X., Huang, J. Z., Deng. S. "FP-Outlier: Frequent Pattern Based Outlier Detection". Computer Science and Information Systems, 2005; 2(1):103-118.
4. Moens. S. Aksehirli, E., Goethals. B., "Frequent Item set Mining for Big Data", IEEE International Conference on Big Data,.2013 ; 111-118
5. Adda. M., Wu. L., Feng. Y., "Rare Item set Mining". In Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA '07) Washington, DC. IEEE Computer Society. 2007; 73–80.
6. Li Y., Chung W., Bae HY. " A Novel Outlier Detection Method for Spatio-Temporal Trajectory Data". In: Lee G., Howard D., Ślęzak D. (eds) Convergence and Hybrid Information Technology. ICHIT 2011. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.2011; 6935
7. Menczer. F., Belew. R., Willuhn. W., "Artificial life applied to adaptive information agents". In Working Notes of the AAAI Symposium on Information Gathering from Distributed, Heterogeneous Databases. AAAI Press.1995.

8. Miguel A. Carreira-Peripinan., "A Review of Dimension Reduction Techinques", A Technical Report, University of Sheffield.
9. Petteri Nurmi, Michael Przybilski, Greger Linden, Patrik Floreen, "An Architecture for Distributed Agent-Based Data Pre processing**",** Springer-Verlag Berlin Heidelberg, V. Gorodetsky, J. Liu, and V.A. Skormin (Eds.): AIS-ADM 2005, LNAI 3505, 2005; 123–133,
10. Charu C. Aggarwal, Philip S. Yu**,** Watson. T. J, "An Effective and Efficient Algorithm for High-Dimensional Outlier Detection", The VLDB Journal — The International Journal on Very Large Data Bases,2005; 14(2): 211-221.
11. Angiulli.F., Pizzuti., "Outlier Mining in Large High Dimensional Data Sets", IEEE Transactions in Knowledge and Data Engineering 2005; 2(17): 203–215,
12. Eskin E., Arnold A., Prerau M., Portnoy L., Stolfo S. "A Geometric Framework for Unsupervised Anomaly Detection". In: Barbará D., Jajodia S. (eds) Applications of Data Mining in Computer Security. Advances in Information Security, Springer 2001;6:77-101.
13. Robert Burbidge, Bernard Buxton, "An Introduction to Support Vector Machines for Data Mining, Keynote young Operational Research", 11th Conference, University of Nottingham, Computer Science Department, University college of London, 2004 ; 3-15.
14. Hwanjo Yu., Jiong Yang., Jiawei Han., "Classifying Large Data Sets Using SVMs with Hierarchical Clusters", Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003; 306-315..
15. Messner.S.F., "Exploring the Consequences of Erratic Data Reporting for Cross-National Research on Homicide". Journal of Quantitative Criminology, 1992 ; 8(2): 155-173,
16. Tong Simon, Koller Daphne, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research.2001; 45-66.
17. Zaki. M.J., Ho. C.T, Large-Scale Parallel Data Mining, eds.Springer.2000.
18. Han.J., Pei.J., Yin.Y., Mao.R., "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", Data Mining and Knowledge Discovery, 2004;8(1):53–87.