J. Basic. Appl. Sci. Res., 5(2)31-38, 2015

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

A Survey on Outliers Detection in Distributed Data Mining for Big Data

P.Ajitha[1], Dr.E.Chandra[2]

[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore
[2]Director, Department of Computer Applications, Dr. SNS Arts and Science College, Coimbatore.

Received: July 24, 2014
Accepted: January 10, 2015

ABSTRACT

Outlier detection in distributed data mining for large and high data had become a necessitated research arena in current divulge of information. This survey discusses the distributed data mining strategies and algorithms that are developed for big data. Reasoning's for evolving Distributed Data Mining and Parallel Data Mining are stronger as propensity on generating larger and inherently distributed data sets that intensifies performance and communication insufficiencies. Even scalable data mining algorithms require higher execution time for current stringent requirements of today's vast applications. So the distributed strategies for outlier detection are reviled here, along with algorithms and methodologies with due concentration specified for accuracy, speed up and execution times.
KEYWORDS: Outliers, Distributed Data Mining, classification.

1. INTRODUCTION

Outlier Detection, is finding abnormal or suspicious activity that do not conform to the norm or the expected behavior. These patterns are of irregular and non-conforming and said to be as outliers. For outlier detection, most of the algorithms assume that data are centralized and resides in the single memory hierarchy. The centralised design is unable to handle the issue of incessant increase in the size, complexity of data sources and in the pervasiveness of distributed data sources [5], in this context Distributed Data Mining [7] is considered for Outlier detection.

Distributed Data Mining is mining of data from distributed sources where the data is fragmented into different data nodes called as local nodes and further transferred to the global node. Fragmentation of the data into various nodes, reduces the transfer time and ability to handle large data sets is possible. As outlier detection task is very time consuming [1], so the distributed and parallel version of implementations for outliers detection are developed.

Outlier Detection problem is time consuming, when the data sets are of large size and from various data sources. To reduce time and increase the size and complexity of the data in outlier detection, distributed strategies are developed for detection of outliers in large and high dimensional data sets. The two main problems in centralised issues are communication bottleneck and privacy data loss where data from one source may be too sensitive and private to reveal to others [2].

Existing Outlier detection methodologies deals with two dimensions or attributes of datasets. Statistical based outlier detections need a model of the data points and the relations to the postulated model is to be known in advance to the users [9]. Knowledge of underlying data is necessary for statistical community. Strategies need to be developed for outlier detection for high dimensional data sets.

1.1 What are Outliers?

Outlier is the data objects that deviates from the normal pattern or behavior where it shows as it generated by the different contrivance [3]. Basically outliers are different from noise, whereas noise is an error or variance but outliers are aberration from the normal data and need to be removed before outlier detection.

Outliers may be induced in the data for variety of reasons, e.g., credit card fraud, terrorist activity or breakdown of a system, but all these may have a common characteristic that may be interesting and useful to the researcher. This interestingness is a key feature for outlier detection.

1.2 Challenges

Challenges that arise in outlier detection for the large data sets are firstly defining Normal Objects and outliers or modeling them with specified boundary is very much difficult and tough [3]. As the boundary that differentiates both are very close and sometimes be normal also. Secondly the Outlier Detection, depends on the application domains it is dealt with. Application Specific Outlier Detection also need to be addressed, as the outlier of one may

***Corresponding Author:** P.Ajitha, Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, (0)984382331,ajitha@y7mail.com,

be the signal of another and differs from domain to domain. For example, outliers in clinical data may very form that of marketing data.

Thirdly, handling noise in the outliers is also very challenging task. It blurs the distinction between normal objects and outliers. To some extent the noise may also hide the outliers and reduce the effectiveness of the outliers. Availability of the labeled data for training or validation models used by outlier detection is usually a major issue. Finally, specifying the degree of the outlier, the unlikelihood of the object being outlier generated by normal mechanism is highly challenging task.

These challenges are not easy to solve. This survey just embodies the common methodologies, algorithms that are available in the literature to address the issues of distributed data mining strategies in outlier detection.

1.3 Related Work

Distributed Strategies for Outlier Detection in very large and high dimensional data sets are discussed in the existing literature. Distance based algorithms [9] discusses the nearest neighbors of the outliers that exists. Parallel bay [10] detects outliers for large datasets in the parallel environment. Density Based algorithm with Local Outlier Factor[10] handles the weights of the outliers and how efficiently it can be detected by calculating the reachability distances of the k neighbor outliers. A novel index based method, by utilizing the data structure INDEX is discussed in Detecting OutLiers PHusing objects into Index(DOLPHIN)[11], with disk resident data sets.

1.4 Our contributions

This survey offers the distributed strategies that exist for detecting outliers in large and high dimensional data sets. Most of the outlier detection techniques focus on specific application domains or in single research area and in centralised domain. Focus of this paper is the distributed strategies and algorithms pertaining to it. Reasons for distributed strategies and the some of the aspects that's concerned or not taken into contemplation are also discussed. The categories of outlier detection in distributed environment and the handling of large and di mensional data sets are uncovered in this survey.

1.5 Organizational Flow of the Article

Section 2 discusses the different facets of outlier detection along with the nature of the data, types of outliers and modes of outliers. Section 3 discusses the applications of Outlier Detection in distributed data mining for large and high dimensional data sets. Section 4 describes the Classification Based Outlier Detection, Section 5 confers Nearest Neighbor Based Outlier Detection algorithms. Section 6 discusses Clustering Based Outlier Detections , Section 7 discusses the Statistics Based Outlier Detection. Section 8 confers Distributed Based Outlier Detection Algorithm and Relative strengths and weaknesses of outlier detection are discussed in Section 9. Section 10 discusses Conclusion and Future Works.

2. Different Facets of Outlier Detection in distributed data mining

Outlier detection in distributed data mining is an important research aspect in current environment. Data sets that needs for distributed environment, different types of outliers, mode of outliers are taken into discussion. This section brings forth the richness in the problem domain and justifies the need for outlier detection.

2.1 Nature of the data

A vital aspect of the outlier detection in distributed data mining is the input data. Input data is collection of data instances referred to as object, pattern, point, vector, entity etc. Each of the data instances contains different set of attributes also referred to as variable, feature, characteristic, field, and dimension. The attributes are of univariate or multivariate or mixed set of attributes or both types.

Nature of data determines the applicability of the domain which may be of categorical data or continuous data. Instead of the actual data, pairwise instances in form of distances (similarity matrix)[28].

2.2 Types of Outliers

Different types of outliers are Global Outlier, Contextual Outlier and Collective Outlier. Global outlier is otherwise called as point outlier, where individual data instance that is different from the rest of data is considered as point outliers. Simply, an Object O_g that deviates from the rest of data set. Example of this application is Intrusion Detection in Computer Networks. To find an fitting measurement of deviation is of the necessitated one..

Contextual Outlier or otherwise called as conditional outlier, if an data instance is an outlier in specific context but not otherwise is called as contextual outlier. Object O_c, it deviates significantly based of selected context. Attributes of data object are of two types, contextual attributes which defines the context in terms of time and

location. For example in spatial data sets, contextual attributes are its longitude and latitude of a location. In time series data, time determining the position of a data in an entire sequence is specified as contextual attributes

Second type of attribute is behavioral attribute, defines the non-contextual features of an instance. For example, average rainfall in the entire world, amount of rainfall at any location is a contextual attribute.

Collection of related data instances are outliers with respect to entire from the rest of the data sets is called as collective outliers. Subset of data objects collectively deviate significantly for whole data sets, even if the individual data sets are not outliers. For example, collective outliers where a group or number of computers sends denial- of-service package. For this type of outliers, background knowledge on the relationship of data such as distance or similarity measure on objects is necessary.

2.3 Modes of Outliers

Basically models of outliers are based on whether the user labeled examples of outliers can be obtained. General Application scenarios of outlier detections are supervised, semi-supervised and unsupervised scenarios [4].

In Supervised methods for some applications, training data with some normal and abnormal data objects are available. The data instances are labeled both for normal and outlier. Here, modeling outlier detection is a classification problem and samples examined are of testing and training data. Modeling of normal objects and reporting that those that are not matching the model is called as outlier. Otherwise, model outliers and treat those not matching the model as outlier. They may be also multiple normal and abnormal classes. Challenges of these supervised methods are imbalanced classes and finding outliers as many as possible, i.e., not mislabeling normal objects as outliers.

In Semi-Supervised, either the training data for normal or abnormal classes will be provided. The data instances are labeled only for very small objects. If some labeled normal objects are available, use labeled examples and the proximate unlabeled objects to train model for normal objects.

In Unsupervised scenario, most of the applications never carry training data. Data instances are never labeled. Here, assume normal objects are clustered into multiple groups and outliers are expected to be far away. Detecting outliers effectively in this method are very difficult. In some intrusion or virus detection normal activities are highly diverse. Unsupervised methods have high false positive rate [3] but still miss many outliers. Supervised methods are very effective as it identifies key resources effectively and contains labeled instances for normal and outliers. Clustering methods are suitable for unsupervised methods to find clusters then outliers. Challenge lies in distinguishing noise from outlier, but far less outliers than normal objects. Solution for this is tackling the outliers directly.

3. Applications of Outlier Detection

Applications of the outliers are credit card fraud detection, i.e unusual credit card purchase, Data Cleaning, Telecom Fraud Segmentation, Customer Segmentation, identification of competitor and emerging business trends in e-commerce[27] and Medical Analysis, Intrusion Detection[3].

In Intrusion detection, detection of malicious or suspicious activity in computer related system. Here, volume of data sets is very high and outlier detection technique should be very effective in dealing with it. Supervised and semi-supervised methods will be preferred as data instances are labeled for normal but for intrusions it is not.

Fraud detection is another application domain, where the outliers are detected for criminal activities that occurs in bank, insurance agencies, Mobile companies, credit card companies , stock market etc. Malicious users may be actual customers or might be posing as customer. Fraud occurs when the users consume resources provided by the company in an unauthorized way. In credit card fraud detection, frauds are reflected in transactional records and correspond to high payments, purchase of items never purchased by the user, or high rate of purchase. Challenge here is unauthorized credit card usage that requires detecting that type of outliers.

4. Classification Based Outlier Detection

In Classification, learn a model or classifier from set of labeled data instances and classify test instance using the learnt model. Two phase classification are operated here. Training phase where classifier is learned using the available labeled training data. Testing Phase classifies a test instance as normal or outlier using classifier.

Both multi-class and one class outlier detection techniques are available. One Class SVMs are one of the classification of outliers. Multiclass availability of outliers are very few. Distributed strategies for multi-class classification need to be developed. Various classification based outlier detection are Neural Networks, Support Vector Machines and Bayesian Networks. In rule based classification, Decision trees[29][30][31] and association rules[32]. Both one-class and multi-class classifier are available for DT and AR. FPOF*(Frequent Pattern Outlier Factor)*[27] algorithm discusses algorithm to detect outlier transactions and identify new interesting outliers.

Algorithms related to distributed strategies are discussed in Decision Trees and are used for anomaly detections and compared to other classification methods, it is of less suspect bile to the curse of dimensionality[23]. The data handled for both discrete and continuous distribution with joint distribution is used here. Split points are found out and a stopping criterion threshold is determined with pruning.

5. Nearest Neighbor Based Outlier Detection

Basic assumption here, is the normal data sets instances occur in dense neighborhoods while outliers occur far from the closest neighbor. Distributed strategies for nearest neighbor based are discussed in the distance based and density based algorithms.

5.1 Distance Based Algorithms

Distance Based Algorithms[9] is discussed for finding outliers that has more than two dimensions/attributes. Cell Based Approach, considers more than two dimensions but the data should be Disk Resident. FindAllOutsM algorithm , finds out whether all data that is considered for outlier detection resident in disk or not. Memory Resident is checked and minimizing the number of page reads or passes over data is possible by initial mapping phase and object pair wise phase. The reading of data points and calculation of object by objects are mapped either with same cell or nearby cell.

5.2 ParallelBay's Algorithm

Bay's algorithm Outlier[10] for each instance of datasets and its closest neighbor so far is tracked and stored. Score lower than cutoff data is removed and predicted not as outlier. Sum of the distances of k neighbors and also the average distance and median distance is calculated for the score function. Parallel Bay's algorithm[10] is where distribution of data is done between processes and each process computes local neighbors and results are send to master nodes. Computation of global neighbors is find by master and top outliers are detected.

5.3 Density Based Outlier Detection

Density Based Outlier Detection[10] of each instance for Local Outlier Factor(LOF) is computed which provides the indication how strongly an instance can be considered as an outlier. *k* distance of outliers, *k* distance of neighborhood of an instance, reachability instance, local reachability, density of an instance and local outlier factor of instance is taken for the calculation of density based outlier detection.

5.4 DOLPHIN (Detecting OutLiers PHusing objects into Index)

DOLPHIN[11] is an novel distance based outliers detection algorithm working on disk resident data sets and the Input Output cost is sequentially the cost of reading the dataset twice with the file that is presented. It uses data structure Index for distance based outliers and finds the temporal cost based on the parameters presented. Differs from Cell based algorithm[9] by an data structure called INDEX with three strategies of outlier detection. Three strategies are selection of policies of objects in main memory, pruning rules and similarity search techniques. So DOLPHIN provides better performance than the cell based approach in terms of performance.

5.5 Nested Loop and Index Based

Nested Loop called as NL and Index Based algorithms[9] computes outliers for each input point and is considered with k nearest neighbors and checking of the distances to see whether they are smaller than the nearest neighbor and included in the loop. Distances computation and comparison are continued for all the data points. Index Based algorithm used R* Tree for reducing the number distance based computations.

6. Clustering Based Outlier Detection

Clustering used to group similar data into clusters. It is an unsupervised method possible through semi-supervised clustering. Basic assumption is normal data belongs to cluster and abnormal data or outliers does not belong to any cluster.

Clustering based algorithms like CLARANS[19] where Clustering LArge Applications based on RANdomised Search for mining spatial based data sets. DBSCAN[20] Density Based Spatial Clustering of Application with Noise discusses spatial application detections and noises. BIRCH[21] Balanced Iterative Reducing and Clustering using Hierarchies discuss outliers in the clustering BIRCH algorithms handles large data sets with complexities. and CURE[22] Clustering Using Representatives ,used data points as a representative and applies hierarchial clustering algorithm with also Random Sampling and Partioning.All these algorithms considers outliers but to the extent it

does not include in the predictions and not interfered with their clustering process. These algorithms are sensitive and related to the clusters detected by the algorithms.

7. Statistics Based Outlier Detection

Statistical based methods, deals with assumptions of data normality. General assumption in statistical based is normal data instances occur in high probability regions of stochastic model while anomalies occur in low probability regions of stochastic model[3]. Effectiveness of statistical methods rely on the factor whether the assumption of data helds real data. Gaussian Distribution is to model the normal data. For each object, estimate the probability of data objects fits the Gaussian distribution If it is very low, data is unlikely generated by the Gaussian model, thus an outlier. Parametric, Non-parametric methods are available for statistical based Outliers detection.

8. Distributed Based Outlier Detection Algorithms

Distance Based, Statistical Based algorithms are also handled in the literature, again in terms of disk resident in statistical based and density and k nearest neighbors on distance based algorithms. Again, existing literature have not considered using any distributed data mining techniques like Data Distribution in Distributed Association Rule Mining.

8.1 Outlier Detection Solving Set

Outlier Detection Solving Set Algorithm[6] discusses the subset of data that include sufficient number of points from data to consider the distances among pairs of solving set algorithm and dataset. Computing the solving set and top *n* outliers and classify each unseen object as outlier or not. The classification is done by computing weight with data and solving set. Outlier Detection Problem(ODP) and Outlier Prediction Problem(OPP) are discussed. The ODP finds the objects with greatest weight and OPP considers the weight with regard to the data and distances k , and equal or greater weight is considered to be outlier or inlier. Solving set satisfies efficiency, smallness, meaningfulness and consistency.

8.2 Partition Based Algorithms

In the partition based algorithms size of datasets are not taken and efficient performance is possible only for small size of outliers. Partition based algorithms has not specified whether partition is horizontal or vertical based, which has impact on the type of data it considers. Cell Based Approach considers disk resident datasets so the size of datasets is limited to the main memory size. Size of dimensionality reduction $k \geq 4$ the performance degradation is possible.

Both NL and Index based algorithms are computationally expensive so a partition based algorithm[8] is developed for pruning the k nearest neighbor that cannot be top n outliers as the distances are small. Partition based algorithms consists of data space and pruned as soon as the for distance based data points on partition outliers determination possibilities are slim. Generating partition, Computing bounds, Identifying candidate partitions containing outliers, computing outliers from data points in candidate partition are the steps for partition based algorithms.

8.3 HillOut

HilOut[12] algorithm is presented for Outlier Detection which scans the input datasets and have solution approximation with low time complexity sort. It also reduces the number of scans and generates outliers with single scan. Efficiently detects top *n* outliers with large and high dimensional datasets.

8.4 GridBot

A nested loop algorithm is designed for outliers detection in high dimensional data sets with data in random order and improves the performance in near linear time. GritBot[13] tool used for finding the data that is dissimilar and providing why it is the surprising information. But this algorithm does not scale well or perform if the datasets are ordered or if it contains independent values.

8.5 DEMAC(Distributed Exploration of Massive Astronomy Catalogues)

Top K outliers are detected for astronomy catalogs with DEMAC[14] (Distributed Exploration of Massive Astronomy Catalogues)system. PCA is used for outlier detection to reduce the size of data while retaining as much as possible datasets. PCA used for checking the deviations from the correlated data. Communication efficient distributed algorithm for outlier detection in comparing with the centralized version.

8.6 RBRP(Recursive Binning and Re-Projection)

RBRP[15] algorithm is a fast algorithm for mining outlier detection specifically for high dimensional data sets. This algorithm is basically partitioning data into bins and use k-means to cluster the data in the bins. It also makes use of NL algorithm for outliers detection by searching in bins, by first searching iteratively and finding the outliers.

8.7 Twin Signal Sensing Method

Even for the running rotors, outlier detection is applicable by detecting outliers based on the signal signatures and finding out the novel outlier based on the range of signature signals available[16]. It generates the false alarm rate and true alarm rates based on the signature signals detection of outliers. This method discussed is simple, linear and efficient.

8.8 DSS and LDSS

Distributed Solving Set(DSS) and LazyDistributed Solving Set algorithms(LDSS) [17] are distributed strategies for outlier detection in large data sets. DSS algorithm had supervisor node and core computation handled simultaneously by other nodes and synchronization of the partial result after completion of the job. DSS algorithm finds out the outliers by iterative selection of outliers. These outliers are selected by finding the weights of the data sets and selecting the outliers which have maximum outbound weight.

Temporal Cost: Cost Computation of the Distributed Solving Set(DSS) algorithm is computed by considering the distance based between two data set objects. Temporal costs of DSS algorithm for the local nodes are computed. And also supervisor node cost is considered by the retrieval of candidate objects among the total distances of local nodes and heaps consisting of outliers.The overall costs are computed by obtaining the summation of the temporal costs of local and supervisor nodes. When the number of local nodes is very large then run time of the algorithms is small.

Transmission Cost: Amount of the data transferred among the local and supervisor nodes also computed by the procedures NodeInit and NodeComp. NodeInit is executed on each local one time and NodeComp is executed on each local node one time as per DSS algorithm.

In LazyDSS the subset of collection nearest neighbors of each candidate node are computed by starting the smallest ones and sending them into the local nodes to the supervisor node. The NodeComp is modified by not only returning to the smallest nearest neighbors but also the distances of the neighbor also returned. For each candidate objects nearest distance is updated with the entries stored in the local nodes during iteration. After all the nodes are processed with the distances received from each local nodes and candidate additional bunch of distances also computed by NodeReq.

8.9 iORCA

Outlier Detection algorithm iOrca[18] is an improvement of *Orca* algorithm (called as *iOrca)* by means of novel indexing scheme. iOrca is an ring computation which allows the cut-off threshold to be exploited efficiently in distributed fashion. Fast Indexing based induction iOrca is compared with Orca[13] and performs better in terms of speed up and efficiency.

8.10 DROUT

Dimensionality **R**eduction/ Feature Extraction for **OUT**lier Detection[25] is an efficient method for feature extraction in outlier detection. Eigenspace is applied for the training set from the randomly sampled dataset. After the application of eigenspace, where relevant features are extracted and transforms the testing set and detection algorithms are applied. Eigenspace regularization mitigates the loss of discriminant information that arrive from feature extraction.In DROUT , eigen value regularization and feature extraction are performed on eight-adjusted scatter matrix instead of normal ones. DROUT is able to gain accuracy for detection methods.Other factors like size, execution time and speed up is not suggested or included by the method of DROUT.

8.11 Parallel Algorithms for GPU

The parallel algorithms for GPU(Graphical Processing Unit)[26] is derived from algorithms Brute Force and Solving Set Algorithm. The GPU nested loop algorithm is used and updates of heaps is done to the data points by inserting all the distances to it. Parallelisation of GPU nested loop algorithm with LOF is also done. GPU solving set algorithm is inserts distances in the block of shared memory in the heap and hierarchically merge heaps and finally new candidates emerge on the merging of heaps in parallel.

9. Relative Strengths and Weaknesses for Outlier Detection in Distributed Data Mining

Based on the application domains , data sets used the types of outliers are used. Basically when the large data sets are used, Principal Component analysis is used for pre-processing the data. If mixed set of attributes are available, the multi-class classification of outlier can be used. Nearest Neighbor and clustering are not suitable if the data sets dimensions are high, as the normal and outliers differentiation is very difficult. For these type of data sets, classification based techniques will be suitable, as labels are available for both normal and outliers.

Parallel Bay's and density based , LOF outliers detection has a good performance in terms of transmission of data and speed up is achieved through few communications. Execution is not calculated and higher dimensionality of data is also need to be discussed and taken into consideration.

The computational complexity of anomaly detection is key aspect. While classification, clustering and statistical based outlier detection have expensive training times but testing time is very fast. This is acceptable as one of the phase provides quicker result. Unsupervised methods are not suitable when the outliers are very bulky[33][34].

9. Conclusion and Future Works

Outlier Detection in Centralised, Parallel and Distributed Environments are briefly discussed in this paper. For big data, the existing algorithms do not scale well other than DSS and LDSS. Here, number of data sets, size, and execution times all are discussed but again the pre-processing of the data has not been handled. To intensify performance of detecting outliers and new algorithm need to be developed. This paper is the coverage of existing outlier detection in distributed data mining for big data.

The future works will be of pre-processing, classification in detecting outliers with specific criteria for improving performance metrics like accuracy, time is to be of prominent ones.

REFERENCES

1. Angiulli.F, Basta.S, Lodi.S, and Sartori.C, 2010.A Distributed Approach to Detect Outliers in very Large Data Sets,Proc. 16th Int'l Euro-Par Conf. Parallel Processing (Euro-Par), pp. 329-340.
2. Joel.W.Branch,Chris Giannella,Boleslaw K Szymanski, Ran Wolff, Hillol Kargupta, 2013. In Network outlier detection in wireless sensor networks. Knowledge Information Systems. 34(1):23-54.
3. Jiawei Han, Micheline Kamber, and Jian Pei , 2012. Data Mining: Concepts and Techniques, Morgan Kaufmann publishers.
4. Hans-Peter Kriegel, Peer Kröger, Arthur Zimek , 2010.Outlier Detection Techniques. SIAM International Conference on Data Mining.
5. Zaki.M.J and Ho.C.T ,2000.Large-Scale Parallel Data Mining, eds.Springer.
6. Angiulli, F., Basta, S., Pizzuti, C. 2006.Distance-based detection and prediction of outliers. IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 2, February 2006, pp 145–160.
7. Kargupta.H and Chan.P, 2000.Advances in Distributed and Parallel Knowledge Discovery, eds. AAAI/MIT Press.
8. Ramaswamy.S, R.Rastogi.R and Shim.K.2000.Efficient Algorithms for mining outliers from large datasets. Proceedings. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 427-438.
9. Knorr.E and Ng.R, 1998.Algorithms for Mining Distance-Based Outliers in Large Datasets. Proceedings. 24th International Conference Very Large Data Bases (VLDB), pp. 392-403.
10. Lozano.E and Acuna.E, 2005.Parallel Algorithms for Distance-Based and Density-Based Outliers. Proceedings. Fifth IEEE Internatinal Conference. Data Mining (ICDM), pp. 729-732.
11. Angiulli.F and F. Fassetti.F,2009.Dolphin: An Efficient Algorithm for Mining Distance-Based Outliers in very Large Datasets. IEEE Transactions Knowledge Discovery from Data, vol. 3, no. 1, article 4.
12. Angiulli.F and Pizzuti.2005.Outlier mining in large high dimensional data sets, IEEE Transactions in Knowledge and Data Engineering 2(17):203–215.
13. Bay.S.D and Schwabacher.M. 2003.Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proceedings Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.
14. H. Dutt., C. Giannella, K. D. Borne, and H. Kargupta.2007. Distributed top-k outlier detection from astronomy catalogs using the demac system. Proceedings SIAM Internatinal Conf. Data Mining .
15. A. Ghoting, S. Parthasarathy, and M. E. Otey.2008.Fast mining of distance-based outliers in high-dimensional datasets. Data Mining Knowledge Discovery.,Vol 16(3):349–364.

16. S. E. Guttormsson, R. J. Marks, M. A. El-Sharkawi, and I. Kerszenbaum. 1999.Elliptical novelty grouping for on-line short-turn detection of excited running rotors. IEEE Transactions on Energy Conversion, 14(1):16–22.
17. Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori, 2013. Distributed Strategies for Mining Outliers in Large Data Sets IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 7.
18. Kanishka Bhaduri, Bryan L. Matthews,2011.Algorithms for speeding up distance-based outlier detection, SIGKDD 11 San Diego, CA, USA
19. Raymond T. Ng and Jiawei.1994. Han. Efficient and effective clustering methods for spatial data mining. In Proceedings of the VLDB Conference, Santiago, Chile.
20. Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu.1995 A database interface for clustering in large spatial databases. In International Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), Montreal, Canada.
21. Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996.Birch: An efficient data clustering method for very large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 103–114, Montreal, Canada.
22. Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. 1998.Cure: An efficient clustering algorithm for large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data.
23. Matthias Reif, Markus Goldstein, Armin Stahl, Thomas M. Breuel, 2008.Anomaly Detection by Combining Decision Trees and Parametric Densities , 19th IEEE International Conference on Pattern Recognition, 2008. ICPR 2008.
24. He, Z., Xu, X., Huang, J. Z., Deng, S.2005. FP-Outlier: Frequent Pattern Based Outlier Detection. Computer Science and Information Systems, Vol. 2, No. 1, 103-118.
25. Hoang Vu Nguyen,Vivekanand Gopalkrishnan.2010.Feature Extraction for Outlier Detection in High-Dimensional Spaces, JMLR: Workshop and Conference Proceedings 10: 66-75 The Fourth Workshop on Feature Selection in Data Mining.
26. Fabrizio Angiulli , Stefano Basta , Stefano Lodi and Claudio Sartori.2013. Fast outlier detection using a GPU,Proceedings of the International Conference on High Performance Computing and Simulation (HPCS), Helsinki, Finland.
27. Agyemang M, Barker K and Ahajj R 2006 A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis 10, 521-538.
28. Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.
29. Fan, W., Miller, M., Stolfo, S. J., Lee, W., and Chan, P. K. 2001. Using artificial anomalies to detect unknown and known network intrusions. In Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society, 123-130.
30. Helmer, G., Wong, J., Honavar, V., and Miller, L. 1998. Intelligent agents for intrusion detection. In Proceedings of IEEE Information Technology Conference. 121-124.
31. Lee, W. and Xiang, D. 2001. Information-theoretic measures for anomaly detection. In Proceedings of the IEEE Symposium on Security and Privacy. IEEE Computer Society, 130.
32. Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering. IEEE Computer Society, Washington, DC, USA, 3-14.
33. Sun, J., Xie, Y., Zhang, H., and Faloutsos, C. 2007. Less is more: Compact matrix representation of large sparse graphs. In Proceedings of 7th SIAM International Conference on DataMining.
34. Soule, A., Salamatian, K., and Taft, N. 2005. Combining filtering and statistical methods for anomaly detection. In IMC '05: Proceedings of the 5th ACM SIGCOMM conference on Internet measurement. ACM, New York, NY, USA, 1-14.