Distance based and Multivariate Parallel Techniques for Outlier Recognition

Dr. C.P.Gupta and Arunima Sharma
Department of Computer Science and Engineering
University College of Engineering
Rajasthan Technical University, Kota, INDIA

Abstract- In Information Age various types of statistics is collected from profuse sources. If we represent this census in graphical form we get deviation in avail. This divergent is high or less than other significance of data set. The exception in the distinguished ability of these values is unambiguously conspicuous, if depicted in graphical form with resulted ethics. Outlier is a dossier of data values which is either greater or lesser means drifted from other resulted esteemed used in representation of a data establish in representable form. Numerous techniques are used to distinguish outliers from other values by using LOF with outlier identification schemes. Outlier detection is deploying on distance measures, clustering and spatial methods. These approaches are used in description and understand the sufficient data of miscellaneous innovation which is used to diagnose and designated outliers from other information set values of homogeneous avails. In this paper we discuss about two such fields of outlier recognition, Distance based and multivariate parallel techniques which are highly used in several applications for outlier detection.

INTRODUCTION

Nowadays data is presented in form of sets that are so large and complex that it becomes difficult to process and sometimes referred as big data. Statistical data values are numerical values which are easy to represent in graphical form. These data includes group of several values and inspect data values in a set of samples. The large values are chronicled and analysed for getting an overview of recorded value in graphical manner because graphs make it easy to represent relation among values on basis of some parameters or factors. There are variations among these value sets which make some data values different from others in the set of values observed (Gao et. al. 2010). These deviated values are known as outliers. Occasionally outliers are appraised as an error in data set but both outlier and noise or errors are non-identical. Noise may misshape the accustomed objects (Ben 2005) and blur the distinction between common objects and outliers. It may help camouflage outliers and reduce the strength of outlier detection.

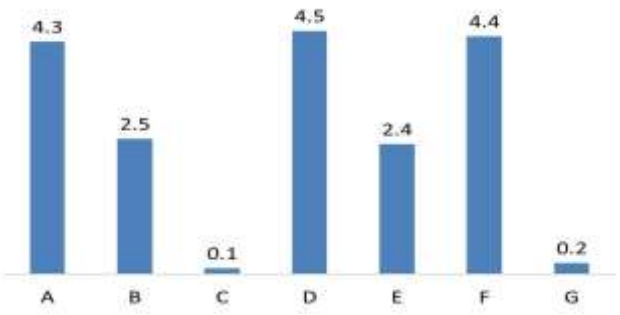

Figure 1- Histogram representation of data set where two values are too low unexpectedly from rest of data values in set 0.1 and 0.2 are considered as outliers.

APPLICATIONS OF OUTLIER DETECTION

Importance of outliers is not bounded in a particular field, Medical area, fraud revealing, clinical prosecutions, voting unevenness analysis (Nguyen 2007), data cleansing, network intrusion (Zanero et. al. 2004), simple weather prediction, geographical information systems (Hodge et. al. 2004), athlete performance scrutiny, and other data-mining tasks are various different field where outlier detection is used. Criminal activities (Zhang et. al. 1995) like unauthorized access considered as fraud (Ngai et. al. 2011) in economic sector like banks, credit card companies, insurance companies, cell phone companies, stock market etc. Companies want to detect these deceitful trades before it do any harm by using outlier identification. Insider trading is also a zone where outlier detection can

be useful where people creating illegal profits by using insider information of companies before it is made public. Medical health anomaly (Li et. al. 2015) detection can be applied to patient chronicles which contains the results of different tests did on the patient. This can lead to encounter of difficulties by patient. Fault in machines like engines, generators, converters etc. or devices in space shuttles are become to identify by outlier detection in their outcomes and working. Generally in huge sized things like images, lots of unpredicted outliers be likely to sneak in Inconsistent data is very diverse from normal data and can be distant though outlier detection only.

CLASSIFICATION OF OUTLIERS

Outliers are categorized (Vinueza et. al. 2004) on origin of referred object arguments in global and local outliers and on basis of result as labelled and scored outliers. Global outliers are conventional set of all values excluding one which create false outcome for remaining set values. Local outlier (Gao et. al. 2006) has not appropriate properties and these encompass trivial subset of data substances. Labelled outliers present outcome of any process in binary format. Scored outliers has constant and scored (established on probability for being an outlier) output.

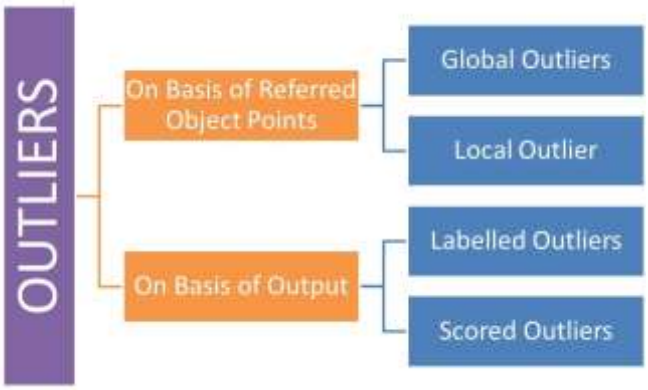

Figure 2 – Classification of outliers

CHALLENGES IN OUTLIER DETECTION

Most difficult task in outlier detection to design a model for its identification, because for each type of data we need different techniques to handle it, and which technique is suitable

based on user input and output desired. The parameters used to classify outlier have to be accurate else data points which are not outliers counted as one of them and decrease result accuracy. Noise is also making difficult to classify outliers in a data set. If clusters (Duan et. al. 2009) are not in a format way then outlier detection become a complex task because it's always based on prior known information. Small data set and normal values (which are less diverse from each other) are highly (Ro et. al. 2015) complex during discovery of outlying values.

Lots of outlier detection methods have been established to resolve focused problems related to a specific application area, while others have been established in a more common manner. A significant challenge in outlier recognition is that it includes discovering the unseen data. Describing a normal province which incorporates usual behaviour is very hard. A lot of times usual behaviour keeps changing and a present perception of common behaviour might not be appropriately illustrative for the future process (Koufakou et. al. 2008). The periphery among normal and outlying comportment is often ambiguous. Thus an outlying value which present near to the borderline can be truly ordinary value and vice versa. Except these common issues Curse of Dimensionality and Sparse data are two major challenges during outlier identification. As the data increase with time rapidly it become difficult to analyse it and it became sparse it is known as curse of dimensionality. And from each point of view of data outlier points vary which make difficult to reach at any conclusion. This type of data is known as sparse data.

PARALLEL ALGORITHMS FOR OUTLIER DETECTION

Parallel algorithms (Oku et. al. 2014) are used in several systems to improve performance and efficiency of device by providing multi-processing to core processor of a system which make device to handle more than one tasks parallel. From 2000's till now several modifications are done in processors, it become multi core from single core processor but by applying frequency scaling during implementation in 2004 single core processor can work equivalent to multi-processor in less time with efficient result. In outlier detection parallel algorithms plays important role. Parallel algorithms provide user to manage Big-data parallel in less time. This High-Dimensional data (Von et. al. 2015) is not only complex and unstructured but also difficult to handle because of velocity of its generation and need of fast outcome.

MULTIVARIATE FUNCTIONAL OUTLIER DETECTION

When more than one parameters means Bivariate and Multivariate (Hubert et. al. 2015) parameters jointly are considered for detection of outlier instead of univariate value for a single variable in data set to identify multivariate outlier which is contradictory to univariate outliers, is known as multivariate functional outlier detection technique (Hong et. al. 2015). This method is based on data structure and size of data known as cut off value which is used to categorized (Wu ett. al. 2013) outliers among several cluster points (Patel et. al. 2011). This cut off value is measured on basis of difference in distance distribution calculated by Mahalanobis distance. It is one fast and easy method to implement for outlier detection.

BAG-DISTANCE ALGORITHM

Multivariate functional data desired to concatenate robustness, affine invariance and computing feasibility of data. Affine Variance is a complex parameter which does not fit in depth based method (Zimek et. al. 2012) because if point is present outside convex hull

(Filzmoser et. al. 2005) value of depth function become zero, which give false results. To overcome this drawback Bag-Distance (bd) is proposed which consider half value of depth function. Bag- Distance is considered as standard algorithm because of appropriate directional result. It assumes that outlier values are located faraway from dense centre of data set. It divides data in subsets where distance among points is less, if points have high depth from centre data point they are considered as outlier. So all outlier has high rank in depth list based on Bag-Distance and easy to identify. If distance between data point and centre is equal to size of data group the process terminates because it is boundary point in set.

The main distinguishing advantage of Bag-distance is that it permits detecting outliers based on their distance to a position which is measured as best central position in the data set. It is combination of affine invariance easiness, common applicability and strength of k-nearest neighbour (Hautamaki et. al. 2004) and workings fine for both multivariate and functional data types. Bag-distance is used where multimodal distributions fail.

One of the shortcomings of bag-distance is that bag-distances of outlying interpretations are neither predominantly large nor small; henceforth do not respond outlyingness properly (Angiulli et. al. 2009).

HRS15 ALGORITHM

The concept of Bag-distance is modified in HRS15 Algorithm. Bag- distance is based on distance to location of data point from centre. Centre point is those who are deepest point in set of values and then it standardized on set of values. HRS 15 is also a systematic method for outlier detection in univariate and multivariate (Filzmoser et. al. 2008) data sets. HRS 15 consist a factor Depth-Outlyingness-Quantile-Rank (D-O-Q-R) which link depth and outlyingness inversely. In HRS15 at a time interval t functioning is based on probability and weight factor of all values in depth based procedure. One of the advantages is that HRS15 can be applied in such applications where calculation of depth is a complex process. It is more accurate for multivariate data because it considers more than one parameter for outlier detection. Weight allow to classify hidden outliers which poses weird behaviour like based on a parameter they are outlier but on basis of other parameter they behave as cluster entity.

Depth is appropriate only for disseminations that are unity-modal with convex. Therefore, for non-convexly sustained distributions or for multimodal allocations, depth is not appropriate for imitating centrality in set. The non-outlying curves does not consist small depth value, because of randomness in data points. Except all these features HRS15 cannot detect the distribution measure used by weight functions.

DISTANCE-BASED OUTLIER DETECTION

In set of random data values where top data values are required as outcome of outlier detection process Distance-based outlier detection methods are most preferable, especially for unlabelled data set. It generates a sub set of values based on distance among data points. Outlierness is calculated on basis of the distance among a pair of top data points required. Among several properties of outlier detection it is useful and unique because of sub-quadratic time required in computing and identification of solution in high-dimensional data (Aggarwal 2015). Distance-based outlier detection is an efficient approach for experimentation on synthetic and real data set. The unusualness among neighbour (Hautamaki et. al. 2004) data

points is identified on basis of distance among them which make it a non-parametric approach for unusualness identification.

ORCA

There are several distance based outlier detection algorithms are there who has main features like fast computation time, non-parametric methodology, unusual distance, and others. Among several different algorithms ORCA is one of best algorithm developed by Bay and Schwabacher, based on Pruning and Random nested looping rules. It is used in numerous fields like network deceit detection, data cleaning, intrusion identification, error detection done by humans, and others (Chauhan et. al. 2015). In ORCA if distance between data object and its adjacent neighbour is small it is considered as cluster value else as outlier. To eliminate data objects and get a set of outlier value cut-off value is used to eliminate other values and get top k values only. According to observations ORCA works in linear time for high-dimensional database. It does not require any pre-processing before analysis (Nguyen et.al. 2006) to detect index value of data, which make it suitable to process dynamic data.

Shortcoming of Orca-based outlier detection algorithm is that computation time for manipulative the outlier significance of a statistics (Xiong et. al. 2005) objects in a record differs with each statistics entity. Every worker thread required to update and read values from outlier-score table which create conflicts in values arise because of shared score table among all worker threads.

DISTANCE-BASED OUTLIER SCORE (DBOSk(o))

Distance-based outlier (Knorr et. al. 2000) score of a data object ‘o’ is the distance of the K^{th} nearest neighbour (Brito et. al. 1997) DBOSk(o). Similarly TDBOSk(o,m) is distance between K^{th} neighbour from object ‘o’ to m^{th} object. ORCA create list of top t outliers based on their distance value and top t outlier values to which are below cut off values. Time complexity of DBOSK(o,m) is O(dN2) where d is dimension and N is total objects in data set. Mostly DBOSK(o) run exponentially and it doesn’t require any indexing.

Shortcomings of DBOSK(o) is that it is a Parametric-based method which has restrictions for several real world applications to find correct representations and identify parameters for classification. Scattering of data entities in DBOSK(o) is not known previously and data objects contain of multivariate data.

DPWH (Data Parallelism with hierarchical Outlier Detection)

To reduce conflict in Outlier score table hierarchy based model DPWH is used which make cache of cut off values for each individual thread. Performance of DPWH is better than DBOSK(o), because it include three phases Data parallelism (Christopher et. al. 2015), Hierarchical Outlier Score Table and Caching of Global Outlier score. In parallelism data is sub divided in partitions which is processed by core CPU’s. It use Round Robin algorithm for equal distribution of data in each partitioned block and balance load and computation time. Each thread manage Hierarchical table for top t outlier and update it periodically. To improve performance of DPWH Speed to identify large outlier entities and detection (Micenková et. al. 2015) of data to be prune are used.

Round-robin partition is superior than block partition algorithm is advantage of DPWH over DBOSK(o), but it depends on total numbers of row in data set. It provides a well-organized

parallelization prototype for parallel data processing like string explorations by using suffix tree and multicore CPU. DPWH is has independent of data level in database.

Like Advantages DPWH have some Shortcomings too. If usage of a thread is completed early when supervision of data assigned to other thread is not it needs to execute until processing of other thread does not over. For block partition based data set DPWH is not a suitable choice for processing.

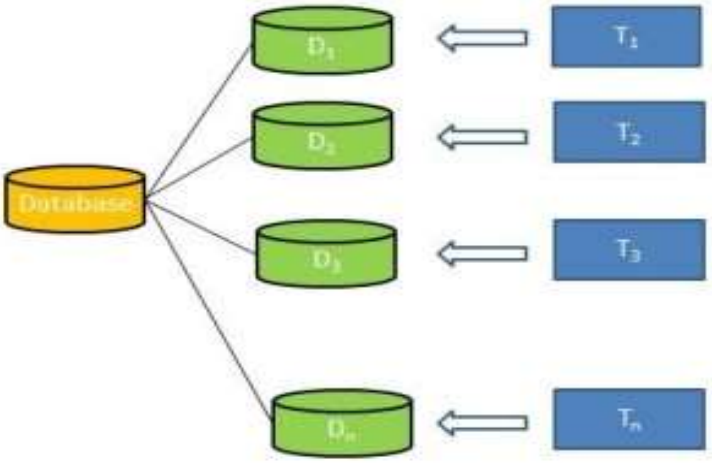

Figure 3 – Database partition in DPWH

SUPERVISOR ALGORITHM

Previous research in field of clustering and outlier detection, mostly are based on distance and does not applicable for high dimensional data (Aggarwal et. al. 2001), which require parallel distributed processing. Supervisor algorithm is one of them which distribute data in sub sets and process them parallel. Data set in supervisor algorithm considered as global data on node which is known as supervisor node. Central processing (Han et. al. 2011) is done by all other nodes and result is send to supervisor node in end. In central processing nodes receive data set, calculate distance among each element of set and their weights too.

An advantage of Supervisor Algorithm is that it as a final point gives the accuracy of results. It decreases data conversions and message cost in data set. The supervisor algorithm runs for distribution of data among all nodes and synchronized partial results returned by each node to others after they complete their task. To calculate the novel lower bound for weight it needs distinct processing. Performance of supervisor algorithm is based on Communication time to transfer data and processing time taken by node. Shortcomings of this technique are that it is not appropriate for huge data sets and for mining (Miller e. al. 2008) of distributed data, because it needs all data to be conveyed among all network nodes.

NODEINIT PROCEDURE ALGORITHM

Data transfer among supervisor node and other node is carried out by NodeInit procedure algorithm. NodeInit procedure executes at each node to identify outlier values in data set. It not only identifies distance among all nodes but also calculate weight at each node. After calculation of weight arrange nodes in ascending order, Find median value of data set which present at centre. If data points are of odd numbers it is easy to identify it else there are two median values present for a set. Mark one quarter points as lower bounder points, which may

be one if data points are even else two in number if data points are odd. As increase in total number of nodes, processing time will decrease and efficiency of system increases. NodeInit procedure is also applicable to modify such algorithms which are based on Support Vector Machines (Erfani et. al. 2016) to their parallel version. As Shortcomings it can be distinguished that the total number of distances calculated is faintly increasing with the number of nodes. The distances figured is little complex with respect to number of nodes in set. Distances and weights are computed with respect to each node are based on the distances among them. Each node has equal load balance for precise data.

DEVIATION BASED OUTLIER DETECTION

In Deviation based methodology outliers are those values that do not possess common local (Yuan et. al. 2014) features of data set and after their exclusion the variation in data set become minimum. Deviation based outlier detection techniques are based on independent distribution of data and they use heuristic search and naïve algorithm concept while data handling.

DENSE SUBGRAPH PARTITION (DSP)

The Dense sub-graph (Liu et. al. 2015, Shekhar et. al. 2001) partition is based on hyper graphs division. DSP created in end are always unique. This algorithm has two phases of partition in first phase it is classified sequentially on basis of vertices and in second it is pseudo disjoint sub-graph partition. The DSP is different on all other algorithms is that it does not consist any parameter value and sub graphs are generated automatically. Sub-graphs obtained from first phase are known as conditional core sub-graphs (CCS). But CCS identification is time consuming, complex and costly process. For partition of graph it use Min-Partitioning algorithm based on divide and conquer and permutation in random order. Min-Partitioning algorithm create significant clusters created in bottom up way in which are time efficient and useful for parallel processing. The classification of graphs (Akoglu et. al. 2015) is based on its area of application according to it is decided which values are unrelated in cluster.

But Shortcomings of this is that there is no partition technique which gratify preferred prerequisite. The main restriction of DSP derives from its explanation of density (Breunig et. al. 2000), which is the ratio among aggregate weight and the number of vertices. Though, the aggregate weight be influenced by the total hyper edges, which grows much quicker than the number of vertices on hyper-graphs. Therefore, DSP and its sub graphs are very large and hence original cluster structure is never identified.

COMPARISION

In outlier detection the depth value of points outside convex hull become zero which make depth function of no use. Bg-Distance Algorithm is used to overcome this problem by dividing whole data space in half parts at a time and extract outliers according to it. Another depth based algorithm is HRS15 which is used for data space which is used in non-convexly distributions (Billor et. al. 2000) and multimodal distributions of data points. HRS15 is used at that place where depth may fail properly reflecting centrality of data points. The complexity of data is increased due to their several copies and it needs large and costly hardware to manage and solve them. Large memory is required to save and operate operations on such large amount of data. Conventional Distance based algorithms (Campos

et. al. 2016) are not efficient so to overcome these drawbacks ORCA is introduced. ORCA is based on nested loop with randomization.

ORCA managing outlier score tables is complex and time consuming operation. ORCA eliminate problem related to several copies of same data value in outlier score table in univariate data set. Therefore by using concept of top t data objects and cut-off threshold with conventional ORCA model new Algorithm DBOSk is developed for industrial purpose it is an improved version of ORCA it use Worker threads to handle Database on multiprocessing CPUs parallel. NodeInit procedure is based on PDM/DDM Approach to handle distributed and parallel data management flexibly. It is specially used for outlier detection in irrelevant data set. Graphical database is analysed for outlier detection by using deviation based approaches (Chandola et. al. 2009).

Graphs represent complex inter-related data which is complex and contain several dimensions. Hyper-graph is one of the complex data structure which represent multidimensional data. For outlier identification we partition hyper-graph and apply DSP algorithm on it. DSP is a complex large algorithm which is based on divide-and-conquer algorithm, called min-partition evolution.

CONCLUSION AND FUTURE WORK

Swiftly growing data s need fast handling within less time which create need of parallel processing for cluster data mining. Data is classified and enhance in principal step of dispensation which differentiate data set in two classes' outliers and non-outliers. Outlier detection plays vital role, it require a predefined method before grouping of data set but it is tough to apply one procedure on data collected from different sources for different purposes based on various limits. In statistics outlier identification is actually an essential and its importance fall with time required identifying them which erect fast waged algorithms. In future we require conceptual and conditional techniques for outlier detection for more accuracy and assurance that our selected value is really an outlier or not.

In Future work we require to create such multi-dimensional algorithms which work on expected probabilistic data values. The deviation analysis in outliers is sensitive with respect to the view of observing angle of values which variants outliers in each direction and generates several values which confuse user in taking any decision. In future work can be done in Feature Selection and Memory constrained Algorithms development. High-Dimensional Data requires large memory space to save data and to apply analysis on it. The operations on data require arranging them base on different values and after that outliers and clusters are created. Each data value in database contains some attributes which are used for its representation. Data having several characteristics make it difficult to represent. Decision regarding any data value is outlier or not, requires two or more fields which increase complexity of process, which arise requirement of basic efficient feature selection strategies.

REFERENCES

Aggarwal, C. C. (2015). *Outlier analysis. In Data mining* (pp. 237-263). Springer International Publishing.

Aggarwal, C. C., & Yu, P. S. (2001, May). *Outlier detection for high dimensional data*. In ACM Sigmod Record (Vol. 30, No. 2, pp. 37-46). ACM.

Akoglu, L., Tong, H., & Koutra, D. (2015). *Graph based anomaly detection and description: a survey*. Data Mining and Knowledge Discovery, 29(3), 626-688.

Angiulli, F., Fassetti, F., & Palopoli, L. (2009). *Detecting outlying properties of exceptional objects*. ACM Transactions on Database Systems (TODS), 34(1), 7.

Ben-Gal, I. (2005*). Outlier detection. Data mining and knowledge discovery* handbook, 131-146.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). *BACON: blocked adaptive computationally efficient outlier nominators*. Computational Statistics & Data Analysis, 34(3), 279-298.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). *LOF: identifying density-based local outliers*. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

Brito, M. R., Chavez, E. L., Quiroz, A. J., & Yukich, J. E. (1997). *Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection*. Statistics & Probability Letters, 35(1), 33-42.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., ... & Houle, M. E. (2016). *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study*. Data Mining and Knowledge Discovery, 30(4), 891-927.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., ... & Houle, M. E. (2016). *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study*. Data Mining and Knowledge Discovery, 30(4), 891-927.

Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly detection: A survey*. ACM computing surveys (CSUR), 41(3), 15.

Chauhan, P., & Shukla, M. (2015, March). *A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm*. In Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in (pp. 580-585). IEEE.

Christopher, T., & Divya, T. (2015, March). *A Study of Clustering Based Algorithm for Outlier Detection in Data streams*. In Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications(pp. 194-197).

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). *Cluster-based outlier detection*. Annals of Operations Research, 168(1), 151-168.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). *High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning*. Pattern Recognition, 58, 121-134.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). *Multivariate outlier detection in exploration geochemistry*. Computers & geosciences, 31(5), 579-587.

Filzmoser, P., Maronna, R., & Werner, M. (2008). *Outlier identification in high dimensions*. Computational Statistics & Data Analysis, 52(3), 1694-1711.

Gao, J., Cheng, H., & Tan, P. N. (2006, April). *Semi-supervised outlier detection*. In Proceedings of the 2006 ACM symposium on Applied computing(pp. 635-636). ACM.

Gao, J., Hu, W., Li, W., Zhang, Z., & Wu, O. (2010, August). *Local outlier detection based on kernel regression*. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 585-588). IEEE.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hautamaki, V., Karkkainen, I., & Franti, P. (2004, August). *Outlier detection using k-nearest neighbour graph*. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 430-433). IEEE.

Hodge, V., & Austin, J. (2004). *A survey of outlier detection methodologies*. Artificial intelligence review, 22(2), 85-126.

Hong, C., & Hauskrecht, M. (2015). *MCODE: Multivariate Conditional Outlier Detection*. arXiv preprint arXiv:1505.04097.

Hubert, M., Rousseeuw, P. J., & Segaert, P. (2015). *Multivariate functional outlier detection*. Statistical Methods & Applications, 24(2), 177-202.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). *Distance-based outliers: algorithms and applications*. The VLDB Journal—The International Journal on Very Large Data Bases, 8(3-4), 237-253.

Koufakou, A., Secretan, J., Reeder, J., Cardona, K., & Georgiopoulos, M. (2008, June). *Fast parallel outlier detection for categorical datasets using MapReduce*. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on (pp. 3298-3304). IEEE.

Li, Y., Nitinawarat, S., Su, Y., & Veeravalli, V. V. (2015, April). *Universal outlier hypothesis testing: Application to anomaly detection*. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on(pp. 5595-5599). IEEE.

Liu, H., Latecki, L. J., & Yan, S. (2015). *Dense subgraph partition of positive hypergraphs*. IEEE transactions on pattern analysis and machine intelligence, 37(3), 541-554.

Micenková, B., McWilliams, B., & Assent, I. (2015). *Learning representations for outlier detection on a budget*. arXiv preprint arXiv:1507.08104.

Miller, H. J. (2008). *Geographic data mining and knowledge discovery*. The handbook of geographic information science, 352-366.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). *The application of data mining techniques in financial fraud detection*: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569.

Nguyen, D. T., Memik, G., & Choudhary, A. (2006, February). *A reconfigurable architecture for network intrusion detection using principal component analysis*. In International Symposium on Field Programmable Gate Arrays: Proceedings of the 2006 ACM/SIGDA 14 th international symposium on Field programmable gate arrays (Vol. 22, No. 24, pp. 235-235).

Nguyen, T. T. (2007). *Outlier Detection: An Approximate Reasoning Approach*. Lecture Notes in Computer Science, 4585, 495.

Oku, J., Tamura, K., & Kitakami, H. (2014, November). *Parallel processing for distance-based outlier detection on a multi-core CPU*. In Computational Intelligence and Applications (IWCIA), 2014 IEEE 7th International Workshop on (pp. 65-70). IEEE.

Patel, V. R., & Mehta, R. G. (2011). *Impact of outlier removal and normalization approach in modified k-means clustering algorithm*. IJCSI International Journal of Computer Science Issues, 8(5), 331-336.

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). *Outlier detection for high-dimensional data*. Biometrika, 102(3), 589-599.

Shekhar, S., Lu, C. T., & Zhang, P. (2001, August). *Detecting graph-based spatial outliers: algorithms and applications (a summary of results)*. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 371-376). ACM.

Vinueza, A., & Grudic, G. (2004). *Unsupervised outlier detection and semi-supervised learning*. Technical Report CU-CS-976-04, University of Colorado at Boulder.

Von Brünken, J., Houle, M. E., & Zimek, A. (2015). *Intrinsic Dimensional Outlier Detection in High-Dimensional Data*. Technical report, National Institute of Informatics, Tokyo.

Wu, S., & Wang, S. (2013). *Information-theoretic outlier detection for large-scale categorical data*. IEEE transactions on knowledge and data engineering, 25(3), 589-602.

Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S. H. (2005, May). *Analysis of breast cancer using data mining & statistical techniques*. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on(pp. 82-87). IEEE.

Yuan, Y., Zhang, Y., Cao, H., & Yao, R. (2014, May). *New local density definition based on minimum hyper sphere for outlier mining algorithm using in industrial databases*. In Control and Decision Conference (2014 CCDC), The 26th Chinese (pp. 5182-5186). IEEE.

Zanero, S., & Savaresi, S. M. (2004, March). *Unsupervised learning techniques for an intrusion detection system*. In Proceedings of the 2004 ACM symposium on Applied computing (pp. 412-419). ACM.

Zhang, Z., Deriche, R., Faugeras, O., & Luong, Q. T. (1995). *A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry*. Artificial intelligence, 78(1-2), 87-119.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). *A survey on unsupervised outlier detection in high-dimensional numerical data*. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5), 363-387.