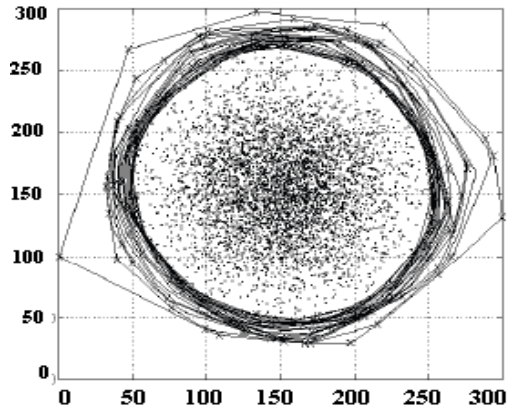

Figure 4: data objects in convex hull layer. [25]

3.2.2 Distance-Based Outlier Detection

Currently, so-called distance-based methods or outlier detection, as typical *non-parametric* methods identify outliers based on the measure of full dimensional distance between a point and its nearest neighbour in the data set.

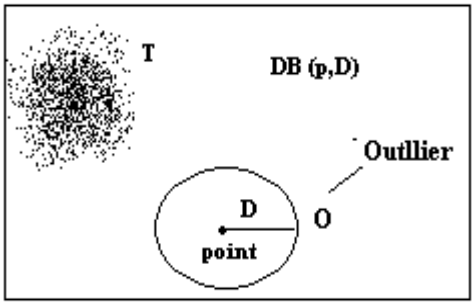

Figure 5: Basic model of distance based method.

The distance-based outlier method was presented in [4], where the definition of outlier becomes: "An object O in a dataset T is a DB (p,D)-outlier if at least fraction p of the objects in T lies at a distance greater than D from O". The parameter *p* is the minimum fraction of objects that must lie outside an outlier's D neighbourhood. Thus the algorithm is further extended based on the distance of a point from its k the nearest neighbour [3]. After ranking points by the distance to its k the nearest neighbour, the top k points are identified as outliers. Alternatively, in the algorithm proposed by Angiulli and Pizzuti [26], the outlier factor of each data point is computed as the sum of distances from its k nearest neighbours. A method for discovering outliers in near linear time has been presented in that randomize the data set for efficient pruning of the search space. Some recent work proposed by Branch et al uses a non-parametric, unsupervised method to detect outliers [27].

3.2.3Deviation Based Outlier Detection

Deviation based approach is used where dataset is having sparsely metric representation. Deviants are outliers defined based on a representation of sparsely metric. The sequential problem approach to deviation-based outlier detection was introduced in Arning et al [5]. These techniques identify outliers by inspecting the characteristics of objects and consider an object as an outlier if the object deviates from these features. Jagadish et al [28] gave the histogram based methods to deal with deviants in time series databases but this method does not fit into data stream scenario. Mining deviants in data stream but the problem of finding an optimal algorithm for deviants in multivariate case was still left open.

3.2.4 Density based Outlier Detection.

Density based outlier detection estimate density distribution of a data point within data set and compares the density around a point with the density around its local neighbour. The relative density of a point compared to its neighbours is computed as an outlier score and points which are having a low density is considered as an outlier.

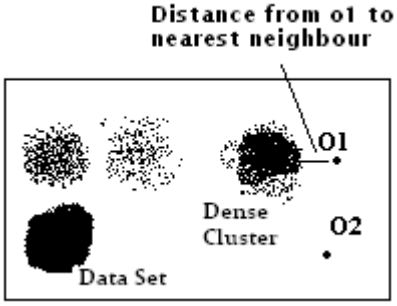

Figure 5: Data set and dense cluster with outliers.

Breunig et al. [29] originally introduce the notion of density-based local outliers based on the density in the local neighbourhood. Each data point is assigned a local outlier factor (LOF) value, which is calculated by the ratio of the local density of this point and the local density of its nearest neighbours. Points that have the largest LOF values are considered as outliers. The LOCI (Local Correlation Integral) method was proposed by Papadimitriouet al [30] which detects outliers based on the metric Multi Granularity Deviation Factor (MDEF) which is a measure of how the neighbourhood count of a particular data element compares with that of the values in its sampling neighbourhood.

3.2.5Clustering Based Outlier Detection.

Cluster analysis is popular unsupervised techniques to group similar data instances into clusters. The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behaviour of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong to a cluster where they are very different from other members. Concept of the cluster-based local outlier proposed by Z. He et al [31], in which a measure for identifying the outlierness of each data object is defined. Wei et al [32] introduced an outlier mining method based on a hyper-graph model to detect outliers in a categorical dataset. The earliest algorithms used or outlier detection are applicable only for single dimensional data sets. Outlier detection for high dimensional data is studied by Aggarwal and Yu [33]. Where data point which lies into low density pattern is called as outlier. Moreover, their algorithm has a high computational cost. The frequent pattern based outliers has been described. A major limitation of clustering-based approaches to outlier detection is that they require multiple passes to process the data set.

3.3 Spatial Outlier Detection Approach

Spatial outliers are spatially referenced objects whose non-spatial attribute values are significantly different from those

of other spatially referenced objects in their spatial neighbourhoods. Spatial outlier detection approaches are broadly categorized into set based and spatial set based.

a) set based outlier detection : Set based outlier detection approach considers statistical relationship among attributes while it ignores spatial relationship among objects. Such approaches are developed for different conditions, type of data distribution, expected outliers and there types.

b) Spatial set based outlier detection: This approach is further classified into space based approach and graph based approach

c) Space based approach*:* Space-based outliers use Euclidean distances to define spatial neighbourhoods. Kou et al. developed spatial weighted outlier detection algorithms which use properties such as canter distance and common border length as weight when comparing non-spatial attributes [34]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbours [35].

d) Graph based approach: Graph based Approach uses graph connectivity to define spatial neighbourhoods. Yufeng Kou et al. proposed a set of graph-based algorithms to identify spatial outliers, which first constructs a graph based on k-nearest neighbour relationship in spatial domain. The algorithms have two major advantages compared with the existing spatial outlier detection methods: accurate in detecting point outliers and capable of identifying region outliers [36].

e) Visualization based approaches: Visualization based approaches try to map the data in a coordinate space and detect the data instances which lie in sparse areas. One of the approaches proposed to detect telecommunications fraud in which abnormal activity appears different on the display and can be visually identified by a user.

4. Advances in Outlier Detection

Traditional Outlier Detection technique may not be applicable to large dataset or categorical data set. Some new techniques proposed for outlier detection are able to detect outliers within high dimensional data or multivariate data.

4.1 Information theory based Approach

Information Theory based techniques analyse the *information content* of a data set using different information theoretic measures such as *entropy, relative entropy etc.* The general idea behind these approaches is that outlying instances affect the information content of the data set because of their surprising nature. Lee and Xiang [37] list different information theoretic measures which can used to detect outliers in a sequence of operating system call. He et al. [38] find a *k*-sized subset from a given data set which when removed makes the entropy of the remaining data set minimal. They use an approximate algorithm called *Local Search Algorithm* (LSA) He et al. [39] to approximately determine this subset of outliers in a linear fashion.

4.2 Support Vector Machine-based Approach

This approach is followed in many areas because of high accuracy & able to handle high dimensional data. Based on the characteristics of the support vectors obtained from SVM-models of varying complexity was proposed. SVM-based methodologies are been widely used for outlier detection, because they do not require a-prior knowledge about any kind of statistical model, can be applied to data with high dimensionality and provide an optimum solution maximizing the margin of decision boundary [40].

4.3 Spectral Decomposition Based Approach.

This approach in general estimates the principal component vector for a given matrix. Using these vectors normal modes of behaviour in the data is detected. Principal component analysis (PCA) is a technique that is used to reduce dimensionality before outlier detection and finds a new subset of dimension which captures the behaviour of the data. PCA based outlier detection approach in wireless sensor network proposed by Chatzigiannakis et al. [41] used to solve to data integrity and accuracy problem. Dutta et al. [42] adopt this approach to detect outliers in astronomy catalogs. Sun et al. [43] propose an outlier detection technique using non-spectral matrix approximation. These techniques are more suitable where data has lot of anomalies and a mixed categorical data.

4.4 Fuzzy logic Based Approach.

This approach uses fuzzy logic for outlier detection. Fuzzy Logic (FL) is linked with the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which object are having degree of membership. Fuzzy Rough Semi-Supervised Outlier Detection proposed by Xue et al. [44]. This approach combines the Semi-Supervised Outlier Detection method, which was proposed by Gao et al. [45], with a clustering method introduced by Huand Yu. [46]. Fuzzy based approach categorized as unsupervised method and supervised methods. Unsupervised based methods are having low performance. To improve performance of outlier detection recently semi supervised outlier detection methods proposed.

4.5 High Dimension-based Approach

High dimensional data contains sparse behaviour finding outlier within such data is difficult problem. According to sparse nature data element projection based ODHDP method is proposed. The basic idea of the approach is to find the outliers by clustering the projections of data set. Aggarwal et al. [47] proposed a new technique for high dimensional outlier detection that finds outliers by observing the density distributions of projections from the data. This new definition considers a point to be an outlier if in some lower-dimensional projection it is located in a local region of abnormally low density.

5. Discussions

Effective outlier detection requires the construction of a model that accurately represents the data. Over the years, a

large number of techniques have been developed for building such models for outlier and anomaly detection. To present effectiveness for outlier detection that require a handle following drawback of existing outlier detection techniques. We provide a review of existing outlier detection scheme with respective data mining and address some weakness that:

a) *Statistical based method*: This method depends upon the data distribution to fit the dataset. Basically statistical methods are applicable for single dimensional data elements. This method is having a curse of dimensionality as dimension increases.
b) Distance based Method: require a distance computation between two data points. If Data is large huge amount of computation required which increases computational cost.
c) *Density based method:* This method requires a prior assumption that the density around a normal data object is similar to the density around its neighbours. The density around an outlier is considerably different to the density around its neighbours. These methods are having a exponential runtime with respect to data dimensionality.
d) Clustering based Method: A major limitation of clustering-based approaches to outlier detection is that they require multiple passes to process the data set.

6. Conclusion

Review of outlier detection technique is proposed with the purpose how traditional methods and recent method work for outlier detection. We conclude from provided review of outlier detection methods is that most existing research focuses on the algorithm based on special background. Efficiency of an outlier detection method depends upon on type of data and data distribution that are processed. Different Outlier detections techniques depend upon different assumption for detecting outliers. For Instance Statistical outlier detection method requires model to fit data that is to be processed, which increase computational cost of outlier detection. Some techniques require a priori knowledge about data distribution in dataset such as distribution based methods. Assumption based method can work quite well if prior assumption made about data is correct.

If no prior information is available about data which is to be processed or property of data changes in unpredictable way with respective time. Over such situation the most efficient solution is to hybrid or combination of many outlier detection techniques having different principle. Such hybrid or combination of techniques will save high computational cost for detecting outliers.

References

[1] Hadi A.S., A.H.M.R. Imon, and M. Werner, ―Detection of outliers, Computational Statistics, vol. 1, 2009, 57-70.
[2] E. M. Knorr and R. T. Ng. ―Algorithms for mining distance based outliers in large datasets In Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pages 392–403, 1998.
[3] Ramaswamy, R. Rastogi, and K. Shim. ―Efficient algorithms for mining outliers from large data sets pages 427–438, 2000.
[4] M. Knorr and R. T. Ng. ―Finding intentional knowledge of distance-based outlier In VLDB ’99: Proceedings of the 25thInternational Conference on Very Large Data Bases, 1999.
[5] Knorr, E. M., Ng, R. T. Algorithms for Mining Distance-Based Outliers in Large Datasets, Proc. 24th VLDB, 1998
[6] M. Ester, H.-P.Kriegel, J. Sander, X. Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, (KDD 96), Portland, Oregon, 1996.
[7] FabrizioAngiulli, Clara Pizzuti, Fast Outlier Detection in High Dimensional Spaces, Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, p.15-26, August 19-23, 2002.
[8] F. Angiulli, S. Basta, and C. Pizzuti.Distance-based detection and prediction of outliers. IEEE Transaction on Knowledge and Data Engineering, 18(2):145(160, February 2006).
[9] M. F. Jiang, S. S. Tseng, C. M. Su. Two-phase clustering process for outliers detection. Pattern Recognition Letters, 2001, 22(6/7)
[10] M.M. Breunig, H.P.Kriegel, R.T. Ng and J.Sander, LOF: Identifying Density-Based Local Outliers ACMSIGMOD 2000.
[11] ArmanKanooni, “A survey of existing data mining techniques, methods and Guidelines within the context of enterprise data warehouse” Master of Science Dissertation in Information Systems.
[12] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In Proceeings of the 29th VLDB conference, 2003
[13] NitinThaper, SudiptoGuha, PiotrIndyk, Nick Koudas, Dynamic multidimensional histograms, Proc. of the 2002 ACM SIGMOD int. conf. on Management.
[14] A. Arning, R. Agrawal, P. Raghavan. A linear method for deviation detection in large databases. In: Proc of KDD’96, 1996
[15] S. Harkins, H. He, G. J.Willams, R. A. Baster.Outlier detection using replicator neural networks. In: Proc of DaWaK’02, 2002
[16] Tian Zhang, Raghu Ramakrishnan, and MironLivny,Birch: An Efficient data clustering method for very large databases, Proc. for the ACM SIGMOD Conf. on Management of Data, Montreal, Canada, June 1996.
[17] M. Ester, H.-P.Kriegel, J. Sander, X. Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, (KDD 96), Portland, Oregon, 1996.
[18] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan.Automatic subspace clustering of high dimensional data for data mining applications. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, Seattle, Washington, June 1998.
[19] K.Prasanna Lakshmi, Dr.C.R.K.Reddy,” A Survey on different trends in Data Streams”, International Conference on Networking and Information Technology 2010.

[20] Yue Zhang, Jie Liu, Hang Li," An Outlier Detection Algorithm based on Clustering Analysis" First International Conference on Pervasive Computing, Signal Processing and Applications 2010.
[21] VitNiennattrakul, Eamonn Keogh, Chotirat Ann Ratanamahatana, "Data Editing Techniques to Allow the Application of Distance-Based Outlier Detection to Streams", IEEE International Conference on Data Mining 2010.
[22] Peng Yang; Biao Huang;" KNN Based Outlier Detection Algorithm in Large Dataset" International Workshop on Education Technology and Training, 2008.
[23] Hawkins D., Identification of Outliers, Chapman and Hill 1980.
[24] Barnett V., Lewis T., Outliers in Statistical Data. John Wiley, 1994.
[25] Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int.Conf.on Knowledge Discovery and Data Mining (KDD), New York, NY.
[26] F. Angiulli and C. Pizzuti, 2002. Fast outlier detection in high dimensional spaces. In Procedings of PKDD'02, 2002.
[27] J. W. Branch et al, 2006, In-network outlier detection in wireless sensor networks, In 26th IEEE International Conference on Distributed Computing Systems (ICDCS'06), pp 49
[28] H. V. Jagadish et al, 1999. Mining Deviants in a Time Series Database. In Proceedings of 25international Conference on Very Large Data Bases. Edinburgh, Scotland, pp 102-113.
[29] M. M. Breunig et al, 2000. LOF: Identifying density-based local outliers. In Proceedings of ACM-SIGMOD Int. Conf.Management of Data (SIGMOD'00).Dallas,TX , pp 93-104.
[30] S. Papadimitriou et al, 2003. LOCI: Fast outlier detection using the local correlation integral. In Proceedings of the 19th International Conference on Data Engineering, Bangalore, India,
[31] Z. He et al, 2003. Discovering cluster based localoutliers. Pattern Recognition Letters.Vol 24, No. 9-10.
[32] Wei et.al, 2002, Outlier detection integrating semantic knowledge. In Proceedings of Third international Conference on Advances in Web-Age information Management. Lecture Notes in Computer Science, Springer- Verlag, London, Vol. 2419.
[33] C. C. Aggarwal and P. S. Yu., 2001.Outlier detection for high dimensional data.In Proc. 2001 ACM-SIGMOD Int.Conf.Management of Data (SIGMOD'01), pp37-46.
[34] Y. Kou, C.-T.Lu and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining, Bethesda, Maryland, USA, 2006.
[35] N. R. Adam, V. P. Janeja, and V. Atluri., "Neighbourhood based detection of anomalies in high-dimensional spatio temporal sensor datasets". In Proceedings of the 2004 ACM symposium on Applied computing, Nicosia, Cyprus, 2004. pp. 576–583.
[36] Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F." Spatial Outlier Detection: A Graph-Based Approach", ICTAI 2007, Volume 1, 2007, pp.281 – 288.
[37] 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.
[38] He, Z., Xu, X., and Deng, S. 2005. An optimization model for outlier detection in categorical data.In Proceedings of International Conference on Intelligent Computing.Vol. 3644. Springer.
[39] He, Z., Deng, S., Xu, X., and Huang, J. Z. 2006. A fast greedy algorithm for outlier mining.In Proceedings of 10th Pacific-Asia Conference on Knowledge and Data Discovery.567-576.
[40] E.M. Jordaan. Deployment of Robust Inferential Sensors, "Irtdusrriol application of Supper Vector Machines for Regression", Ph. D. thesis.Eindhaven University of Technology, 2002.
[41] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B.Maglariset, Hierarchical Anomaly Detection in Distributed Large-ScaleSensor Networks, Proc. ISCC, 2006.
[42] Dutta, H., Giannella, C., Borne, K., and Kargupta, H. 2007. Distributed top-k outlier detection in astronomy catalogs using the demac system.In Proceedings of 7th SIAM International Conference on Data Mining.
[43] 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 Data Mining.
[44] Xue, Z, Shang, Y., Feg, S., Semi-supervised outlier detection based on fuzzy rough C-means clustering, Mathematics and Computers in Simulation, 80, 2010, (pp.2011-2021).
[45] Gao, J.,Cheng, H., Tan, P.N.,. Semi-supervised outlier detection, Proc. of the 2006 ACM Symposium on Applied Computing, ACM Press, 2006, pp. 635–636.
[46] Hu, Q, Yu, D. An improved clustering algorithm for information granulation, in: Proceeding of 2nd International Conference onFuzzy Systems and Knowledge Discovery (FSKD'05), vol. 3613, LNCS, Springer-Verlag, Berlin Heidelberg Changsha,China, 2005, pp. 494–504.
[47] Charu C. Aggarwal and Philip S. Yu. 2005. An effective and efficient algorithm for high-dimensional outlier detection. VLDB Journal, 14: 211-221, Springer- Verlag Publisher.

Author Profile

Mr. Abhishek B. Mankar received Bachelor's Degree in Computer Science Engineering from SGB Amravati University & Pursuing Master Degree in CSE from P. R. Pote (Patil) College of Engineering. Amravati-444602 Maharashtra, India

Prof. Namrata D.Ghuse, Received the Master Degree in Computer Science from SGB Amravati University. Working as a Assistant Professor In Department of Computer Science and Engineering at P. R. Pote (Patil) College of Engineering.Amravati-444602, Maharashtra, India.