International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

A Survey on Different Unsupervised Techniques to Detect Outliers

Shruti S. Rakhe[1], Archana S. Vaidya [2]

[1] *ME 2nd year Student, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, Maharashtra, India*

[2] *Assistant Professor, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, Maharashtra, India*

Abstract - ***In data mining outlier detection refers to the recognition of data point which does not follow the expected pattern or behavior in a particular dataset or is significantly different from other points in a data. In this paper we will review some of the outlier detection techniques and discuss their advantages and disadvantages with respect to various aspects. Outlier detection techniques can be classified into three modes namely unsupervised, semi-supervised and supervised. But, unsupervised outlier detection methods can be further classified as distance based or density based. Many outlier detection techniques are proposed till date. These proposed techniques can be broadly categorized as distribution based (statistical), clustering-based, density-based and model-based approaches***

Key Words: *Outlier detection, distribution based, clustering-based, density-based and model-based approaches.*

1. INTRODUCTION

Outlier detection is very important part of data mining, whose aim is to detect extraordinary values in a given dataset. An outlier can have several definitions, based on statistics and computer science. Some of the definitions are: a) "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"[1] b) outliers is an "observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [2]. or simply c) "an unexpected event or entity"[3]

Outlier detection has various applications in numerous fields. Some of them are: detecting fraudulent applications for credit cards, detecting deceptive usage of credit cards or mobile phones, to detect fraudulent applications or potentially problematical customers, detecting unauthorized access in computer networks, detecting mobile phone fraud by monitoring phone usage or suspicious trades in the equity markets, monitoring the performance of computer networks, for example to detect network bottlenecks, to detect mislabeled data in a training data set, In satellite image analysis outlier detection can identify novel features, misclassified features or identifying unfamiliarity in images - for surveillance systems or robot neotaxis, Medical condition monitoring - such as heart-rate monitors, motion segmentation - detecting image features moving independently of the background, to identify novel molecular structures in Pharmaceutical research, to detect the onset of news stories, for topic detection and tracking or for traders to determine equity, FX trading stories, commodities, outperforming or underperforming commodities, Detecting unexpected entries in databases which ultimately detect errors for data mining, frauds or valid but unexpected entries, Due to so many applications precise detection of outliers becomes must. Many outlier detection methods are suggested till date. We will categorize and review some of the existing methods in the following sections. The overview of the techniques which will be discussed is given in the figure below:

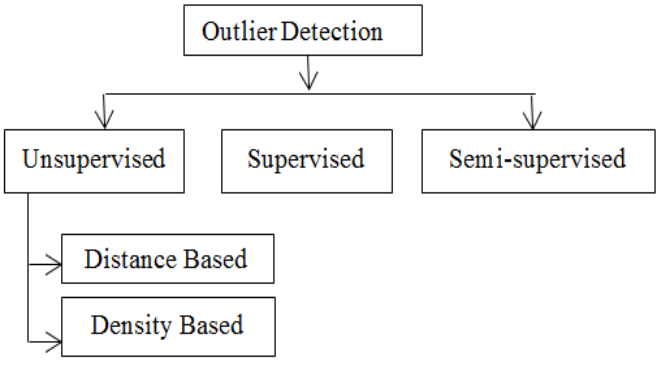

Fig -1: Modes of operation of outlier detection techniques

2. OUTLIER DETECTION TECHNIQUES

There are three fundamental modes of operation to the problem of outlier detection:

1) Unsupervised: It is the process in which no information about the dataset's class distribution is available beforehand. This approach is widely used now a day. We will discuss techniques using this mode of operation later in this paper.

2) Supervised: In this process, dataset consists of class objects that are already classified as normal or abnormal. The work described in [4] uses FMN algorithm for outlier detection which is an example of supervised learning approach. But the limitation of FMN method is that, user has to tune the parameters to get good recognition accuracy. The recognition accuracy at the cost of recall time is increased in the above stated method.

3) Semi-supervised: This approach needs pre-classified data but only learns data which is marked normal. The normal class is taught but the algorithm learns to recognize abnormality. It can learn the model gradually as new data arrives, tuning the model to improve the fit as each new epitome becomes available. It aims to define a boundary of normality. The work in [5] uses SSODPU algorithm which is semi supervised. It deals with the problem of detecting outliers with only few labeled positive examples. There are two main steps in this algorithm: Initially some of the reliable negative examples will be extracted using KNN. And the second step is the fuzzy clustering including both positive and negative example of outlier. Here outliers are detected on the basis of new labeled examples. The limitation of this method is that accuracy is not up to the mark.

Now the techniques for unsupervised outlier detection can be classified into two categories: Distance based and density based. We will see them on by one.

2.1 Distance based approaches

In distance-based approaches, the distances between an object and its nearest neighbors are determined, and then used to estimate the outlierness of an object. Basically the distance-based approaches assume that outliers are far apart from their neighbor objects [6]. Any appropriate distance measure can be used, such as Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. Usually, the type of the variables affects the choice of distance measure. Several well-known methods based on this idea are discussed here.

S. Chawla and A. Gionis in [7] present a technique using which we can simultaneously cluster and discover outliers in data. This approach is the generalization of K-means approach and hence it is NP-Hard. It is an iterative approach and it converges to local optima. This algorithm is not suitable for all similarity measures. But, number of outliers cannot be determined automatically.

In [8] a general framework for handling the three major classes of distance-based outliers inclusive of the long established distance threshold based and the nearest-neighbor-based definitions in streaming environments is proposed. Two novel optimization principles to achieve scalable outlier detection are proposed, and those are "minimal probing" together with "lifespan-aware prioritization". This method is proven to be superlative for determining the outlier status of data points. But modern distributed multi-core clusters of machines are not used to its full advantage to improve scalability.

The work proposed by Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic [9] shows the role of hubness in high dimensional data. They provide AntiHub method using reverse nearest neighbor counts for outlier detection. This method can efficiently find outliers in high dimensional dat. But accuracy may be sacrificed to obtain efficiency sometimes.

In [10] distance based unsupervised method for outlier detection is proposed. It uses iterative random sampling. This method takes inspiration from the simple notion that outliers are not as easily selected as inliers in blind random sampling. Therefore selected objects are given more inlierness scores. A new measure called observability factor is developed using this idea. Moreover entropy of scores is proposed to provide heuristic guideline to find the best size of the nearest neighborhood. But performance of this method deteriorates for highest entropy values. But overall it finds outliers effectively and can be used with the combination of other methods for better results.

Orca is one of the most successful algorithm for the improvement of distance based outlier detection. It is based on nested loop with randomization and a simple pruning rule. Orca-based outlier detection on a multi-core CPU is proposed in [11]. Data parallelism and a multi-thread model is utilized in the proposed parallelization model. Here outlier score tables and cutoff values are shared for pruning among worker threads. Cache of the cutoff value on each worker thread is made and outlier-score tables are managed hierarchically. The proposed model cannot work well for block partition, but it worked well for round-robin partition.

In [12] a novel perspective on clustering high dimensional data is provided. Here instead of trying to avoid the curse of dimensionality, dimensionality is embraced. It is shown that for high-dimensional data clustering hubness is a good measure of point centrality. This paper states that hubs can be used effectively as cluster prototypes. GHPKM method is proposed which proves to be better than K-means. This method provides better inter cluster separation in high dimensional data.

The major drawback of this system is, it only detects hyperspherical clusters, just as K-Means.

An approach to decompose the original tick data matrix by clustering their attributes using a new clustering algorithm Storage-Optimizing Hierarchical Agglomerative Clustering (SOHAC) is proposed in [13]. The proposed approach is established on the grounds that the "pattern of change'' in the tick data remains stable, which allows SOHAC to detect clusters over the entire tick-data matrix. But in long term this "pattern of change'' may vary and may require to update the clustering scheme of SOHAC which will split the original matrix into partitions, each of which can be effectively represented by a single clustering scheme. The detection of such partition in a dynamic way, i.e., as new data are arriving, is also not done in the proposed system.

2.2 Density Based approaches

Distance-based approaches are known to face the local density problem created by the various degrees of cluster density that exist in a dataset. In order to solve the problem, density-based approaches have been proposed. The basic idea of density-based approaches is that the density around an outlier remarkably varies from that around its neighbors [14]. The density of an object's neighborhood is corelated with that of its neighbor's neighborhood. If there is a significant difference between the densities, the object can be considered as an outlier. To implement this idea, several outlier detection methods have been developed recently. The detection methods estimate the density around an object in different ways. Breunig et al. [15] developed the local outlier factor (LOF), which is amongst the most commonly a used method in outlier detection. LOF is influenced by variations like local correlation integral (LOCI)[16],Local distance based outlier factor(LDOF) [17], and local outlier probabilities(LoOP)[18]. Now we will review some density based outlier detection techniques:

[19] proposes an approach for selecting meaningful feature subspace and conducting anomaly detection in the corresponding subspace projection. This approach aims to maintain the detection accuracy in high-dimensional circumstances. The suggested approach determines the angle between all pairs of two lines for one specific anomaly candidate: the first line is connected by the pertinent data point and the center of its adjacent points; the other line is one of the axis-parallel lines. Those dimensions which have a comparatively small angle with the first line are then chosen to constitute the axis-parallel subspace for the candidate. Then, a normalized Mahalanobis distance is introduced to measure the local outlierness of an object in the subspace projection. The proposed algorithm does not deal with nonlinear systems.
An Intrusion Detection System (IDS) is a software application or device that oversees the system or activities of network for policy violations or vicious activities and creates reports to the management system. A number of systems may try to prevent an intrusion attempt but this is neither essential nor awaited for a monitoring system. The main focus of Intrusion detection and prevention systems (IDPS) is to pinpoint the probable instances, logging information about them and in report attempts. Various methods can be used to discover intrusions but each one is specific to a specific method. An intrusion detection system aims to detect the attacks efficiently. An approach to detect the intrusions in the computer network is suggested in [21]. The performance of proposed IDS is better than that of other existing machine learning approaches and almost all anomaly data in the computer network can be significantly detected. The proposed work cannot be used for various distance computation function between the trained model and testing data.

In [22] an outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning is proposed. To deal with data with imperfect labels, likelihood values for each input data are introduced which denote the degree of membership of an example concerning the normal and abnormal classes respectively. The proposed approach works in two steps. In the first step, a pseudo training dataset by computing likelihood values of each example based on its local behavior is generated. Kernel k-means clustering method and kernel LOF-based method to compute the likelihood values are presented. In the second step, the generated likelihood values and limited anomalous examples are incorporated into SVDD-based learning framework to build a more precise classifier for global outlier detection. By integrating local and global outlier detection, proposed method explicitly handles data with imperfect labels and enhances the performance of outlier detection.

Many outlier methods are proposed till date; these existing methods can be broadly classified as: distribution (statistical)-based, clustering-based, density-based and model-based approaches [23]. Statistical approaches [24] assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which don't follow such distributions. The methods in this category always assume the typical example follow a particular data distribution. Nevertheless, we cannot always have this kind of priori data distribution information in practice, mainly for high dimensional real data sets [23].

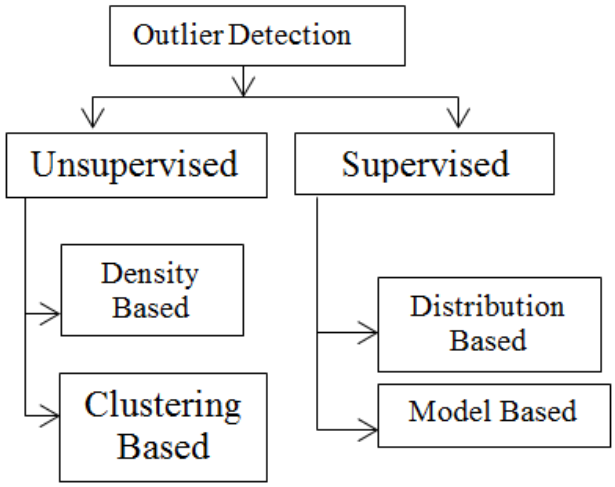

Fig -1: Classification of Outlier Detection techniques based on the availability of training dataset

For clustering-based approaches [25], they always conduct clustering-based techniques on the samples of data to characterize the local data behaviour. In general, the sub-clusters contain significantly less data points than other clusters, are considered as outliers. In the work of [25], the clustering techniques iteratively detect outliers to multidimensional data analysis in subspace. As clustering based methods are unsupervised without requiring any labeled training data, the performance of unsupervised outlier detection is limited.

In [30] a cluster based method for outlier detection is presented. A new global outlier factor and a new local outlier factor and an efficient outlier detection algorithm are developed. This method is can be used with the traditional distance-based outlier detection methods. But it is suggested to use this method as a compliment with other methods of outlier detection.

In [31] two algorithms namely Distance-Based outlier detection and Cluster-Based outlier algorithm for identifying and eradicating outliers using a outlier score are proposed. By cleaning the dataset and clustering based on similarity, one can remove outliers on the key attribute subset rather than on the full dimensional attributes of dataset. This work suggests (based on the results obtained) that cluster based approaches produce better accuracy as compared to distance based methods.

In addition, density-based approaches [15], [26] have been proposed. One of the representatives of this kind of approaches are local outlier factor (LOF) and variants [15]. Based on the local density of each data instance, the LOF calculates the degree of outlierness, which provides suspicious ranking scores for all samples. The most noteworthy feature of the LOF is the ability to estimate local data structure via density estimation. The benefit of these approaches is that they do not demand any assumption for the generative distribution of the data. But, these approaches experience a high computational complexity in the testing phase, since they have to calculate the distance between each test instance and all the other instances to compute nearest neighbors.

In addition to the above contributions, model based outlier detection approaches have been proposed [27], [28], [29]. Among them, support vector data description (SVDD) [27], [28] has been demonstrated empirically to be capable of detecting outliers in various domains. SVDD conducts a small sphere around the normal data and utilizes the constructed sphere to detect an unknown sample as normal or outlier. The most interesting property of SVDD is that it can transform the original data into a feature space via kernel function and effectively detect global outliers for high-dimensional data. However, its performance is affected by the noise involved in the input data.

On the basis of availability of a training dataset, outlier detection techniques described above function in two different modes: supervised and unsupervised modes. Among the four types of outlier detection approaches, distribution-based approaches and model based approaches come under the category of supervised outlier detection, as they assumes the availability of a training dataset that has labeled instances for normal class (as well as anomaly class sometimes).

3. CONCLUSIONS

Outlier detection is very important and has applications in wide variety of fields. So it becomes important to learn how to detect outliers. The main objective of this paper is to review various outlier detection techniques and to study how the techniques are categorized. So we can conclude that, methods used for outlier detection are application specific. Moreover selection of outlier detection method also depends on the type of data involved. In general many authors have suggested that we cannot say that one particular method of outlier detection is the best method. But outlier detection can be efficient if one method is used as a compliment to other method, so that the drawbacks of one method are conquered by use of other method.

ACKNOWLEDGEMENT

We would like to acknowledge Gokhale Education Society's R. H. Sapat College of Engineering, Computer department of their support, guidance and encouragement for writing this paper.

REFERENCES

[1] D. Hawkins. Identification of outliers. Chapman & Hall, London, 1980.

[2] K. Ord. Outliers in statistical data: V. barnett and t. lewis, 1994, 3rd edition, (john wiley & sons, chichester), 584 pp., [uk pound]55.00, isbn 0-471-93094-6. *International Journal of Forecasting,* 12(1):175-176, 1996.

[3] S. Chawla, D. Hand, and V. Dhar. Outlier detection special issue. *Data Min. Knowl.Discov*., 20(2):189-190, 2010.

[4] Nilam Upasania, Hari Omb, "Evolving fuzzy min-max neural network for outlier detection" in International Conference on Advanced Computing Technologies and Applications (ICACTA-2015) *Elsevier*

[5] Armin Daneshpazhouh and Ashkan Sami, "Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering." *IEEE* (2013) 5th *conference on information and Knowledge Technology (IKT)*

[6] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, *ACM Comput.41* (2009)15.

[7] S. Chawla and A. Gionis, "k-means–: A unified approach to clustering and outlier detection." in *SDM. SIAM*, 2013, pp. 189–197.

[8] Lei Cao, Di Yangt, Qingyang Wang, Yanwei Yu+, Jiayuan Wang, Elke A. Rundensteiner, "Scalable Distance-Based Outlier Detection over High-Volume Data Streams" *in ICDE Conference 2014 IEEE*

[9] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovi, "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection" in *IEEE Transactions On Knowledge And Data Engineering*, Vol. 27, No. 5, May 2015

[10] JihyunHa, SeulgiSeok, Jong-SeokLee, "A precise ranking method for outlier detection" in *Information Sciences* 324(2015)88–107 2015 Elsevier.

[11] Junki Oku, Keiichi Tamura, Hajime Kitakami, "Parallel Processing for Distance-Based Outlier Detection on a Multi-core CPU" in 2014 *IEEE 7th International Workshop on Computational Intelligence and Applications* November 7-8, 2014, Hiroshima, Japan.

[12] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic, "The Role of Hubness in Clustering High-Dimensional Data" in *IEEE Transactions On Knowledge And Data Engineering,* Vol. 26, No. 3, March 2014.

[13] Krisztian Buza a, Gábor I. Nagy a, Alexandros Nanopoulos, "Storage-optimizing clustering algorithms for high-dimensional tick data" in Expert Systems with Applications 41 (2014) 4148–4157 2014 *Elsevier.*

[14] V.Chandola,A. Banerjee, V.Kumar, Anomaly Detection: a survey. *ACM Comput.*41(2009)15

[15] M. M. Breunig, H. P. Kriegel, R .T. Ng, J. Sander, LOF: identifying density-based local outliers, in: W. Chen, J. F. Naughton, P. A. Bernstein (Eds.), *Proceedings of the 26th International Conference on ACM SIGMOD, Dallas, USA,* 2000, pp.93–104.

[16] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proc 19th IEEE Int. Conf. Data Eng.*, 2003, pp. 315–326.

[17] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Proc 13th Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2009, pp. 813–822.

[18] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "LoOP: Local outlier probabilities," in *Proc 18th ACM Conf. Inform. Knowl. Manage*., 2009, pp. 1649–1652.

[19] Liangwei Zhang n, Jing Lin, Ramin Karim, "An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection." *in Reliability Engineering and System Safety* 142 (2015) 482-497 *Elsevier.*

[20] Jayanta K. Dutta, Bonny Banerjee, Member, IEEE, and Chandan K. Reddy, Senior Member, IEEE, "RODS: Rarity based Outlier Detection in a Sparse Coding Framework" in *IEEE Transactions on Knowledge and Data Engineering,* Volume: PP Issue: 99 September 2015.

[21] Jabez J, Dr.B.Muthukumar, "Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach" in *International Conference on Intelligent Computing, Communication & Convergence (*ICCC-2015*) Elsevier.*

[22] Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao, and Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels" in *IEEE Transactions On Knowledge And Data Engineering,* Vol. 26, NO. 7, JULY 2014.

[23] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, "Anomaly detection via online over-sampling principal component analysis," *IEEE Trans. Knowl. Data Eng*., vol. 25, no. 7, pp. 1460–1470, May 2012.

[24] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," *Knowl. Inform. Syst.,* vol. 26, no. 2, pp. 309–336, 2011.

[25] Y. Shi and L. Zhang, "COID: A cluster-outlier iterative detection approach to multi-dimensional data analysis," *Knowl. Inform. Syst*., vol. 28, no. 3, pp. 709–733, 2011.

[26] K. Bhaduri, B. L. Matthews, and C. Giannella, "Algorithms for speeding up distance-based outlier detection," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, 2011, pp. 859–867.

[27] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Mach. Learn.,* vol. 54, no. 1, pp. 45–66, 2004.

[28] [28] D. M. J. Tax, A. Ypma, and R. P. W. Duin, "Support vector data description applied to machine vibration analysis," in *Proc. ASCI,* 1999, pp. 398–405.

[29] E. M. Jordaan and G. F. Smits, "Robust outlier detection using SVM regression," in Proc. *IJCNN*, 2004, pp. 1098–1105.

[30] Xiaochun Wang, Xia Li Wang, Yongqiang Ma, D. Mitchell Wilkes, "A fastMST-inspired kNN-based outlier detection method" in *Information Systems48(2015)89–112 Elsevier.*

[31] Christy.A, Meera Gandhi.G, S. Vaithyasubramanian, "Cluster Based Outlier Detection Algorithm For Healthcare Data" in 2nd *International Symposium on Big Data and Cloud Computing* (ISBCC'15)) *Elsevier* 2015.