IPHV8I2024X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

RANKING OUTLIER DETECTION FOR HIGH DIMENSIONAL DATA USING SYMMETRIC NEIGHBORHOOD RELATIONSHIP

[#1]P.RAJASHEKAR, M.Tech Student,
[#2]P.BALAKISHAN, Associate Professor,
Department of CSE
JYOTHISHMATHI INSTITUTE OF TECHNOLOGICAL SCIENCES,KARIMNAGAR,T.S,INDIA.

ABSTRCT: Outlier Detection in high dimensional information turns into a rising system in today's examination in the region of information mining. It tries to discover elements that are significantly disconnected, exceptional and conflicting as for the normal information in a data database. It faces different difficulties as a result of the expansion of dimensionality. Hubness has as of late been produced as a vital idea and goes about as a trademark for the expansion of dimensionality associating with closest neighbors. Grouping likewise demonstrates an imperative part in taking care of high dimensional information and a critical device for exception recognition. This paper builds up a method where the idea of hubness, particularly the antihub (focuses with low hubness) calculation is inserted in the resultant groups got from bunching systems, for example, K-implies and Fuzzy C Means (FCM) to identify the anomalies mostly to lessen the calculation time. It analyzes the consequences of the considerable number of procedures by applying it on three distinctive genuine information sets. The Experimental results show that when every one of the three calculations are looked at, KCAnthub gives a noteworthy decrease in computational time than Antihub and FCAntihub. It is presumed that when the Antihub is connected into K-implies, it beats well.

Keywords: *Outlier, K-NN, High dimensional dataset, Hubness, Antihub*

I. INTRODUCTION

In spite of the huge measure of information being gathered in numerous exploratory and business applications, specific occasions of hobbies are still very uncommon. These uncommon occasions, regularly called exceptions or irregularities, are characterized as occasions that happen occasionally (their recurrence ranges from 5% to under 0.01% relying upon the application). Discovery of exceptions (uncommon occasions) has as of late picked up a great deal of consideration in numerous areas, extending from video observation and interruption identification to fake exchanges and coordinate advertising. For instance, in video observation applications, video directions that speak to suspicious and/or unlawful exercises (e.g. recognizable proof of movement violators out and about, discovery of suspicious exercises in the region of articles) speak to just a little divide of all video directions. Thus, in the system interruption discovery area, the quantity of digital assaults on the system is regularly a little portion of the aggregate system movement. In spite of the fact that exceptions (uncommon occasions) are by definition rare, in each of these illustrations, their significance is entirely high contrasted with different occasions, making their identification critical. Information digging strategies produced for this issue depend on both managed and unsupervised learning. Regulated learning routines commonly assemble an expectation model for uncommon occasions in light of named information (the preparation set), and utilize it to arrange every occasion [1-2]. The significant disadvantages of regulated information mining strategies include: (1) need to have marked information, which can be to a great degree tedious for genuine applications, and (2) powerlessness to identify new sorts of uncommon occasions. Interestingly, unsupervised learning systems commonly don't require marked information and distinguish exceptions as information focuses that are altogether different from the typical (greater part) information in light of some measure [3]. These strategies are ordinarily called exception/ irregularity recognition procedures, and their prosperity relies on upon the decision of closeness measures, highlight choice and weighting, and so on. They have the upside of distinguishing new sorts of uncommon occasions as deviations from typical conduct, yet then again they experience the ill effects of a conceivable high rate of false positives, basically since already concealed (yet ordinary) information can be likewise perceived as exceptions/ oddities. Regularly, information in numerous uncommon occasions applications (e.g. system movement observing, video observation, web use logs) arrives persistently at a tremendous pace in this way representing a noteworthy test to break down it [36]. In such cases, it is imperative to settle on choices rapidly and precisely. On the off chance that there is a sudden or startling change in the current conduct, it is fundamental to distinguish this change as quickly as time permits. Expect, for instance, there is a PC in the neighborhood that uses just set number of administrations (e.g., Web activity, telnet, ftp) through comparing ports. Every one of these administrations relate to specific sorts of conduct in system activity information. On the off chance that the PC all of a sudden begins to use another administration (e.g., ssh), this will

positively resemble another sort of conduct in system activity information. Henceforth, it will be attractive to identify such conduct when it shows up particularly since it might frequently relate to unlawful or nosy occasions. Indeed, even for the situation when this particular change in conduct is a bit much nosy or suspicious, it is imperative for a security examiner to comprehend the system activity and to redesign the idea of the typical conduct. Further, on-line recognition of irregular conduct and occasions additionally assumes a huge part in video and picture examination [4-6]. Robotized distinguishing proof of suspicious conduct and protests (e.g., individuals crossing the border around secured ranges, leaving unattended baggage at the air terminal establishments, autos driving bizarrely moderate or abnormally quick or with irregular directions) in light of data separated from video streams is at present a dynamic examination region. Other potential applications incorporate activity control and reconnaissance of business and private structures. These undertakings are described by the requirement for realtime handling (such that any suspicious movement can be recognized preceding making damage to individuals, offices and establishments) and by element, non-stationary and frequently uproarious environment. Consequently, there is need for incremental exception recognition that can adjust to novel conduct and give auspicious recognizable proof of abnormal occasions.

II. RELATED WORK

The beginning stage for our examinations is a field where the presence of centers has been all around archived and set up, specifically, Music Information Retrieval (MIR). One of the focal ideas in MIR is that of music similitude. Appropriate demonstrating of music similitude is at the heart of numerous applications including the programmed association and handling of music information bases. In Aucouturier and Pachet (2004), center point tunes were characterized as melodies which seem to be, as indicated by a sound similitude capacity, like a lot of different tunes and in this manner continue showing up unwontedly regularly in proposal records, keeping different tunes from being suggested by any stretch of the imagination. Such tunes that don't show up in any suggestion list have been termed 'vagrants'. Comparable perceptions about false encouraging points in music suggestion that are not perceptually important have been made somewhere else (Pampalk et al., 2003; Flexer et al., 2010; Karydis et al., 2010). The presence of the center issue has likewise been accounted for music proposal in light of community oriented sifting rather than sound substance investigation (Celma, 2008). Comparative impacts have been seen in picture (Doddington et al., 1998; Hicklin et al., 2005) and content recovery (Radovanovic et al., 2010), making this wonder a "general issue in mixed media recovery and suggestion. In the MIR writing, Berenzweig (2007) initially suspected an association between the center issue and the high dimensionality of the component space. The center point issue was seen as an immediate consequence of the scourge of dimensionality (Bellman, 1961), a term that alludes to various difficulties identified with the high dimensionality of information spaces. Radovanovic et al. (2010) could give more knowledge " by connecting the center issue to the property of focus (François et al., 2007) which happens as a characteristic result of high dimensionality. Focus is the astounding normal for all focuses in a high dimensional space to be at just about the same separation to every single other point in that space. It is typically measured as a proportion between some measure of spread and extent. For instance, the proportion between the standard deviation of all separations to a self-assertive reference point and the mean of these separations. On the off chance that this proportion unites to zero as the dimensionality goes to unendingness, the separations are said to think. For instance, on account of the Euclidean separation and developing dimensionality, the standard deviation of separations focalizes to a steady while the mean continues developing. In this manner the proportion meets to zero and the separations are said to think. The impact of separation focus has been contemplated for Euclidean spaces and other ℓ p standards (Aggarwal et al., 2001; François et al., 2007). Radovanovic et al. (2010) exhibited the contention that " in the limited case, because of this wonder a few focuses are required to be closer to the information set mean than different focuses and are at the same.

III. HIGH-DIMENSIONAL OUTLIER DETECTION

The high-dimensional case is particularly challenging for outlier detection. This is because, in high dimensionality, the data becomes sparse, and all pairs of data points become almost equidistant from one another [22, 215]. From a density perspective, all regions become almost equally sparse in full dimensionality. Therefore, it is no longer meaningful to talk in terms of extreme value deviations based on the distances in full dimensionality. The reason for this behavior is that many dimensions may be very noisy, and they may show similar pairwise behavior in terms of the addition of the dimension-specific distances. The sparsity behavior in high dimensionality makes all points look very similar to one another. A salient observation is that the true outliers may only be discovered by examining the distribution of the data in a lower dimensional local subspace [4]. In such cases, outliers are often hidden in the unusual local behavior of lower dimensional subspaces, and this deviant behavior is masked by full dimensional analysis. Therefore, it may often be fruitful to explicitly search for the appropriate subspaces, where the outliers may be found.

IPHV8I2024X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

This approach is a generalization of both (full-dimensional) clustering and (full data) regression analysis. It combines local data pattern analysis with subspace analysis in order to mine the significant outliers. This can be a huge challenge, because the simultaneous discovery of relevant data localities and subspaces in high dimensionality can be computationally very difficult. Typically evolutionary heuristics such as genetic algorithms can be very useful in exploring the large number of underlying subspaces [4]. High-dimensional methods provide an interesting direction for intensional understanding of outlier analysis, when the subspaces are described in terms of the original attributes. In such cases, the output of the algorithms provide specific combinations of attributes along with data locality, which resulted in such data points being declared as outliers. This kind of interpretability is very useful, when a small number of interesting attributes need to be selected from a large number of possibilities for outlier analysis.

IV. META-ALGORITHMS FOR OUTLIER ANALYSIS

In many data mining problems such as clustering and classification, a variety of meta-algorithms are used in order to improve the robustness of the underlying solutions. For example, in the case of the classification problem, a variety of ensemble methods such as bagging, boosting and stacking are used in order to improve the robustness of the classification [146]. Similarly, in the case of clustering, ensemble methods are often used in order to improve the quality of the clustering [20]. Therefore, it is natural to ask whether such meta-algorithms also exist for the outlier detection problem. The answer is in the affirmative, though the work on meta-algorithms for outlier detection is often quite scattered in the literature, and in comparison to other problems such as classification, not as well formalized. In some cases such as sequential ensembles, the corresponding techniques are often repeatedly used in the context of specific techniques, though are not formally recognized as general purpose meta-algorithms which can be used in order to improve outlier detection algorithms. The different meta-algorithms for outlier detection will be discussed in the following subsections. There are two primary kinds of ensembles, which can be used in order to improve the quality of outlier detection algorithms: In sequential ensembles, a given algorithm or set of algorithms are applied sequentially, so that future applications of the algorithms are impacted by previous applications, in terms of either modifications of the base data for analysis or in terms of the specific choices of the algorithms. The final result is either a weighted combination of, or the final result of the last application of an outlier analysis algorithm. For example, in the context of the classification problem, boosting methods may be considered examples of sequential ensembles. In independent ensembles, different algorithms, or different instantiations of the same algorithm are applied to either the complete data or portions of the data. The choices made about the data and algorithms applied are independent of the results obtained from these different algorithmic executions. The results from the different algorithm executions are combined together in order to obtain more robust outliers.

A. Neighborhoods

The neighborhood is defined as the set of points lying near the object and thus affecting its anomaly score. There are two types of neighborhoods; the kneighborhood and the r-neighborhood.

These neighborhoods are explained below.

k-distance(p) is equal to d(p,q) where q ∈D and q satisfies the following conditions. The 5-distance(p) is shown in figure 3.1(a).

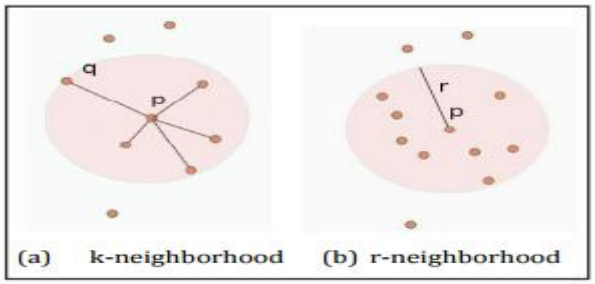
1. For at least k objects q' ∈ D it holds that d(p,q0) ≤ d(p,q)
2. For at most k-1 objects q' ∈ D it holds that d(p,q0) < d(p,q) k-neighborhood(p) is the set of objects that lie within k-distance(p).

The shaded region in fig. 1(a) shows the k-neighborhood.

r-neighborhood(p) is the set of objects lying within r distance from p. The shaded region in fig. 1(b) shows the r-neighborhood.

The k-neighborhood(p) would be denoted by Nk (p) and the r-neighborhood by N(p,r) for the rest of the thesis.

Density based approaches that use k-neighborhood can face some problems in case there are duplicates in the data set. This arises as the density is inversely proportional to the distance and in case we have at least k+1 duplicates of some point then the k-distance would be equal to 0 and thus the estimated density would be infinite. The solution that was proposed in [3] was utilized for these cases. The solution states that the conditions of the k-distance defined above would only apply to objects with distinct spatial coordinates. Meaning that if we have D = {p1,p2,p3,p4} where the coordinates of p2 is the same as p3 and d(p1,p2) = d(p1,p3) ≤ d(p1,p4) , then 2-distance(p1) would correspond to d(p1, p4) and not d(p1,p3). It should be noted that the k-distance(p) is always unique, while the cardinality of the k-neighborhood set could be greater than k.



for k=5. Fig. 1: Neighborhoods Examples

IPHV8I2024X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

V. PROPOSED SYSTEM

Proposed system uses the semi-supervised method which is used half training data. It gives more accurate result as compared to the unsupervised method. The Proposed methodology for outlier detection is explained in this section. In the previous work, unsupervised distance based method used for outlier detection. In the Proposed method semi supervised distance based outlier detection method is used. The advantage of this method is, it gives more accurate result as compared to the unsupervised distance based method. The method is implemented with four phases.

1. In the first phase, import the data set.
2. In the second phase preprocess the data set. Here unsupervised learning approach is used. And calculation of Antihub using the entropy of objects.
3. Outlier detection results.

VI. KNN: K NEAREST NEIGHBORS

K nearest neighbors is a global distance based algorithm. The neighborhood used for this algorithm is the k-neighborhood. The anomaly score is either set to the average distance of the nearest k neighbors similar to the algorithm proposed in [2] or to the k-distance like the algorithm proposed in [2]. The earlier approach has equation

$$knn(p) = \frac{\sum_{o \in N_k(p)} d(p,o)}{|N_k(p)|}$$

LOF: Local Outlier Factor

Local outlier factor was originally proposed in [3]. This is the first local density based algorithm. LOF uses the k-neighborhood. Local density based methods compare the local density of the object to that of its neighbors. For the LOF to accomplish that the following definitions were used.

reach-dist(p,o) The reachability distance is the maximum of d(p,o) and k-distance(o). It is mainly introduced for smoothing local density.

Local reachability density (lrd) The local reachability density of object p relative to Nk(p) is the inverse of the mean reachability distance over the neighborhood set.

Local Outlier Factor (LOF) The local outlier factor is the ratio between the average local reachability density of the neighborhood to that of the object

$$LOF_{N_k(p)}(p) = \frac{\sum_{o \in N_k(p)} lrd_{N_k(o)}(o)}{|N_k(p)| \cdot lrd_{N_k(p)}(p)}$$

The values of the LOF oscillate with the change in the size of the neighborhood. Therefore to improve the results a range is defined for the size of the neighborhood and the maximum LOF score over that range is taken as the final score. The authors of [3] provided some guidelines for choosing the bounds of the neighborhood size range. The lower bound should be greater than 9 in order to smooth statistical fluctuations and it should represent the size of the smallest non-outlying cluster that can be present in the data set. The upper bound should represent the maximum number of objects that can possibly be local outliers which is typically around 20. Normal data would have a LOF score of approximately equal to 1, while outliers will have scores greater than 1. This is explained by the fact that if the data lies within a cluster then local density would be similar to that of its neighbors getting a score equal to 1. For a sufficiently large data set a LOF score of up to 2 would indicate that the point is normal. As the local density based methods are able to detect outliers that were unseen by the global methods and because of the easy interpretability of its score several variants of LOF were developed. Some of which are explained in the following sections.

VII.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 Kmeans 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 nearestneighbor-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 multithread 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 outlierscore 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 highdimensional 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, densitybased 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].

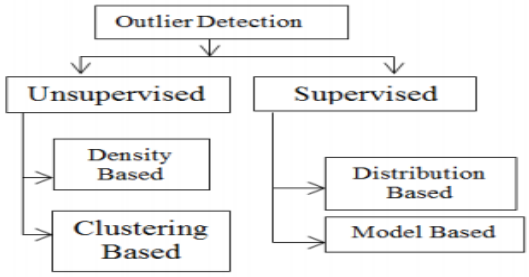


Fig -1: Classification of Outlier Detection techniques based on the availability of training dataset.

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).

VIII. CONCLUSION

The distance based supervised unsupervised etc. approaches used for the outlier detection over high dimensional datasets. A Different technique uses the different concepts such hubness, antihub sets to detect the outliers. Outlier scores also play an important role in outlier detection. This Paper presents a detailed survey of literature which was carried out ia data set for outlier detection. Based on the literature a new approach is proposed i.e. semi supervised learning method for outlier detection. 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.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.