 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

The Competent Reverse Nearest Neighbors for Outlier Detection in High Dimensional Data

[1]Dora Pavani, [2]K Rajendraprasad
[1,2]Dept. of CSE, Miracle Educational Society Group of Institutions, AP, India

Abstract
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. "Hubness" has as of late been recognized as a general issue of high dimensional information spaces, showing itself in the rise of articles, alleged center points, which have a tendency to be among the k closest neighbors of an extensive number of information things. As an outcome numerous closest neighbor relations out there space are hilter kilter, that is, article y is amongst the closest neighbors of x yet not the other way around. The work exhibited here talks about two classes of techniques that attempt to symmetrize closest neighbor relations and explores to what degree they can relieve the negative impacts of center points. We assess neighborhood separation scaling and propose a worldwide variation which has the upside of being anything but difficult to rough for huge information sets and of having a probabilistic understanding. Both neighborhood and worldwide methodologies are appeared to be viable particularly for high-dimensional information sets, which are influenced by high hubness. Both routines lead to a solid diminishing of hubness in these information sets, while in the meantime enhancing properties like characterization exactness. We assess the strategies on a substantial number of open machine learning information sets and manufactured information. At last we show a certifiable application where we can accomplish essentially higher recovery quality.

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 sam

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. 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.

V. Nearest-Neighbor Based

Algorithms that are based on nearest-neighbor based methods assume that the outliers lie in sparse neighborhoods and that they are distant from their nearest neighbors.

Through out the remainder of the thesis let k denote a positive integer, ra real number, D the data set and partitions $\{o,p,q\} \subseteq D$.

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).

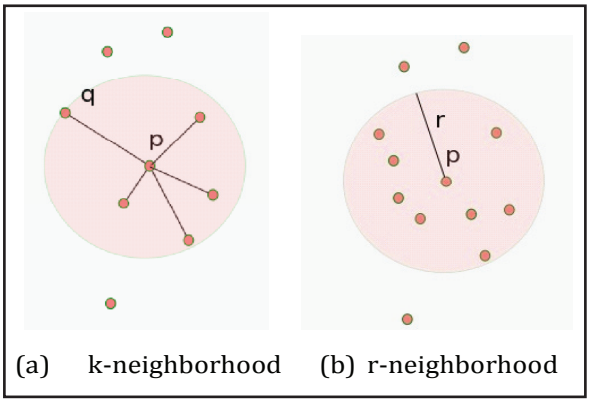
1. For at least k objects q' ∈ D it holds that $d(p,q^o) \leq d(p,q)$
2. For at most k-1 objects q' ∈ D it holds that $d(p,q^o) < d(p,q)$

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 $N_k(p)$ 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



(a) k-neighborhood (b) r-neighborhood

for k=5.

Fig. 1: Neighborhoods Examples

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 $D = \{p_1,p_2,p_3,p_4\}$ where the coordinates of p_2 is the same as p_3 and $d(p_1,p_2) = d(p_1,p_3) \leq d(p_1,p_4)$, then 2-distance(p_1) would correspond to $d(p_1, p_4)$ and not $d(p_1,p_3)$.

It should be noted that the k-distance(p) is always unique, while the cardinality of the k-neighborhood set could be greater that k.

VI. 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 semisupervised 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.

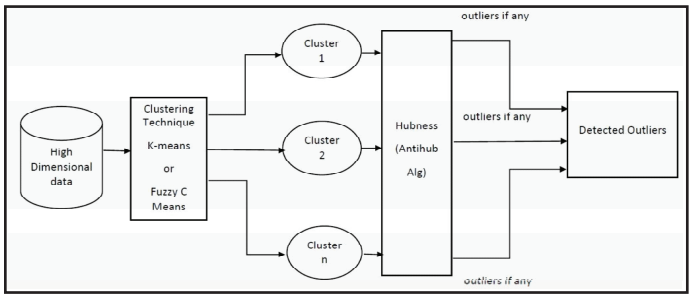


Fig. 2: Proposed System

VII. 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.

$|N_k(p)|lrd^{Nk(p)}(p) = {}^{P}o \in N_k(p)$ reach-dist(p,o)

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 oscillates 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.

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.

References

[1] W. Lee, S. Stolfo,"Data mining approaches for intrusion detection", Proc. of the 7th USENIX security symposium, 1998.

[2] E. Bloedorn, et al.,"Data Mining for Network Intrusion Detection: How to Get Started", MITRE Technical Report, August 2001.

[3] A.K. Jones, R.S. Sielken,"Computer System Intrusion Detection: A Survey. Technical report", University of Virginia Computer Science Department, 1999.

[4] F. Angiulli, S. Basta, C. Pizzuti,"Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 2, pp. 145-160, 2006.

[5] E.M. Knox, R.T. Ng,"Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. Int"l Conf. Very Large Data Bases, 1998.

[6] M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander,"LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int"l Conf. Management of Data, 2000.

[7] W. Jin, A.K.H. Tung, J. Han, W. Wang,"Ranking Outliers Using Symmetric Neighborhood Relationship," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2006.

[8] Yuh-Jye Lee, Yi-RenYeh, Yu-Chiang Frank Wang, "Anomaly Detection via Online Oversampling Principal Component Analysis," IEEE Trans. Knowledge and Data Eng., Vol. 25, No. 7, July 2013.

[9] Y.-R. Yeh, Z.-Y. Lee, Y.-J. Lee,"Anomaly Detection via Oversampling Principal Component Analysis," Proc. First KES Int"lSymp. Intelligent Decision Technologies, pp. 449-458, 2009.

[10] AmolGhoting, Srinivasan Parthasarathy, Matthew Eric Otey,"Fast Mining of Distance-Based Outliers in High-Dimensional Datasets", Springer, 2008.

D. Pavani Holds a B.Tech certificate in Information Technology Engineering from the University of JNTU Kakinada. She presently Pursuing M.Tech (CSE) department of Computer Science Engineering from Miracle Educational Society Group of Institutions, Vizianagaram. Her area of interest includes Data mining and other advances in Computer Applications.

K.RAJENDRA PRASAD is an Assoc. Professor in the Department of Computer Science and Engineering in Miracle Educational Society Group of Institutions, He awarded his M.Tech [Ph.D] degree from JNTU Kakinada. There are a few of publications both national and International Conferences Journals to his credit. His area of interest includes Data mining, Cloud Computing and other advances in Computer Applications.