ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Novel Approach of Usefulness of Reverse Neighbour Counts in Unsupervised Outlier Detection

P. Munwar Gouse[1], B.Nagalakshmi[2], K.Ramesh. [3]

M.Tech Student, Dept. of CSE, GATES Institute of Tech, Gooty, Ananthapuramu, India[1]

Associate Professor, Dept. of CSE, GATES Institute of Tech, Gooty, Ananthapuramu, India [2]

HOD, Dept. of CSE, GATES Institute of Tech, Gooty, Ananthapuramu, India [3]

ABSTRACT: Outlier detection is an approach of finding outlying prototype from the given dataset. Outlier detection grew to become predominant field in specific talents domains. Information size is getting doubled each year there's a have to observe outliers in huge datasets as early as viable. In high-dimensional knowledge outlier detection presents quite a lot of challenges given that of curse of dimensionality. By using examining again the inspiration of reverse nearest neighbors in the unsupervised outlier-detection context, high dimensionality can have a further have an influence on. In excessive dimensions it was once located that the distribution of facets in reverse-neighbor counts turns into skewed .This proposed work ambitions at developing and comparing one of the vital unsupervised outlier detection ways and endorse a strategy to make stronger them. This proposed work goes in details in regards to the development and analysis of outlier detection algorithms such as neighborhood Outlier aspect (LOF), neighborhood Distance-headquartered Outlier component(LDOF), Influenced Outliers and .The principles of those methods are then combined to put in force a new approach with distributed strategy which improves the outcome of the prior recounted ones on the subject of speed, complexity and accuracy.

KEYWORDS: Outlier detection, excessive-dimensional information, reverse nearest neighbors, unsupervised outlier detection methods.

I. INTRODUCTION

1) Detection of outliers in knowledge outlined as discovering patterns in knowledge that don't conform to ordinary habits or data that don't conformed to anticipated habits, this sort of data are called as outliers, anomalies, exceptions. Anomaly and Outlier have an identical which means. The analysts have powerful interest in outliers for the reason that they are going to represent significant and actionable know-how in various domains, such as intrusion detection, fraud detection, and medical and wellness analysis. An Outlier is an observation in information occasions which is special from the others in dataset. There are numerous causes due to outliers arise like bad information pleasant, malfunctioning of apparatus, ex bank card fraud.

Information Labels associated with knowledge occasions shows whether that illustration belongs to usual data or anomalous. Situated on the supply of labels for data example, the anomaly detection approaches function in one of the most three mode's are 1) Supervised Anomaly Detection, techniques educated in supervised mode recollect that the supply of labeled instances for ordinary as well as anomaly classes in a training dataset.
2) Semi-supervised Anomaly Detection, systems proficient in supervised mode keep in mind that the supply of labeled occasions for usual, don't require labels for the ambiguity category. 3) Unsupervised Anomaly Detection, approaches that operate in unsupervised mode do not require training knowledge.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

There are more than a few methods for outlier detection founded on nearest neighbors, which consider that outliers appear a ways from their nearest neighbors. Such approaches base on a distance or similarity measure to go looking the neighbors, with Euclidean distance. Many neighbor-based ways incorporate defining the outlier ranking of a point as the gap to its kth nearest neighbor (okay-NN process), some methods that examine the rating of a factor in step with its relative density, on the grounds that the distance to the kth nearest neighbor for a given knowledge point can also be seen as an estimate of the inverse density around it.

II. RELATED WORK

Assign an anomaly rating known as local Outlier component (LOF) to a given knowledge example. For any given knowledge example, the LOF ranking is equal to ratio of traditional nearby density of the k nearest neighbors of the illustration and the regional density of the data example itself.

To find the local density for a data instance, the authors first find the radius of the smallest hyper-sphere centered at the data instance that contains its *k* nearest neighbors. The local density is then computed by dividing *k* by the volume of this hyper-sphere. For a normal instance in a dense region, there local density will be similar to that of its neighbors, if its local density will be lower than that of its nearest neighbors , then it is an anomalous instance,. Hence the anomalous instance will get a higher LOF score. In [3] Author propose outlier detection approach, named Local Distance-based Outlier Factor (LDOF), which used to detect outliers in scattered datasets. On this to measure how so much objects deviate from their scattered local. Makes use of the relative distance from an object to its neighbors. The greater violation in degree of an object has, the most often object is an outlier. In [4] proposed on a symmetric local relationship measure considers each neighbors and reverse neighbors of an object when estimating its density distribution .To restrict quandary, when outliers are within the region the place the density distributions within the regional are drastically exceptional.

In [5] writer recommends a data circulate outlier detection algorithm SODRNN centered on reverse nearest neighbor. Maintain the sliding window model, to become aware of anomalies outlier queries are carried out so as in the current window. Improves efficiency with the aid of replace of insertion or deletion best in a single scan of the present window.

In [6] advocate a outlier rating centered on the objects deviation in a suite of principal subspace projections. It excludes irrelevant projections displaying no clear change between outliers and the residual objects and to find objects deviating in multiple important subspaces, tackle the general challenges of detecting outliers hidden in subspaces of the info. In [7]writer recommend a unification of outlier ratings furnished via quite a lot of outlier models and a translation of the arbitrary outlier factors to values within the range [0,1] interpretable as values describing the probability of a data object of being an outlier. In [8] advise a new process for parameter-free outlier detection algorithm to compute Ordered Distance change Outlier factor. Formulate a brand new outlier ranking for each illustration by using when you consider that the change of ordered distances. Then, use this worth to compute an outlier score.

III. EXISTING SYSTEMS

A. LOCAL OUTLIER FACTOR (LOF):

In LOF, compare the local density of a instances with the densities of its neighborhood instances and then assign anomaly score to given data instance. For any data instance to be normal not as an outlier, LOF score equal to ratio of average local density of k nearest neighbor of instance and local density of data instance itself. To find local density for data instance, find radius of small hyper sphere centered at the data instance. The local density for instances is computed by dividing volume of k, i.e k nearest neighbor and volume of hyper sphere. In this assign a degree to each object to being an outlier known as local outlier factor. Depends on the degree it determines how the object is isolated with respect to surrounding neighborhood. The instances lying in dense region are normal instances, if their local

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

density is similar to their neighbors, the instances are outlier if there local density lower than its nearest neighbor.LOF is more reliable with top-n manner. Hence it is called as top-n LOF means instances with highest LOF values consider as outliers.

B. LOCAL DISTANCE BASED OUTLIER FACTOR(LDOF):

Local distance based outlier factor Measure the objects outlierness in scattered datasets. In this uses the relative location of an object to its neighbors to determine the object deviation degree from its neighborhood instances. In this scattered neighborhood is considered. Higher deviation in degree data instance has, more likely data instance as an outlier. In this algorithm calculates the local distance based outlier factor for each object and then sort and ranks the n objects having highest LDOF value. The first n objects with highest LDOF values are consider as an outlier.

C. INFLUENCED OUTLIERNESS (INFLO):

This algorithm considers the situations when outliers are within the area where neighborhood density distributions are significantly special, for instance, in the case of objects almost a denser cluster from a sparse cluster, this can give improper outcomes. This algorithm considers the symmetric regional relationship. On this for the reason that affect house and when estimating its density distribution also considers each neighbors and reverse neighbors of an object. Assign each and every object in a database a influenced outlierness degree. The bigger inflo implies that the thing is an outlier.

D. DISADVANTAGES:

1. Threshold value is used to distinguish outliers from usual object and cut down outlierness threshold worth will result in high false negative price for outlier detection.
2. Problem arises when knowledge illustration is placed between two clusters, the inter distance between the item of k nearest neighborhood increases when the denominator worth raises leads to high false confident expense
3. Needs to strengthen to compute outlier detection velocity.
4. Needs to support the effectivity of density based outlier detection.

IV. **PROPOSED SYSTEM**

A. DESCRIPTION OF THE PROPOSED SYSTEM:

An input of collection of large data set will be provided to the proposed system, as data is collected from standard data set repositories, data preprocessing will be applied before passing data to the next phase of the system. Further, this preprocessed input is being passed through to the partition module, where these datasets are been partitioned among many nodes from that one of the node is supervisor node and generate partition statistics and this statistical data is been visualized. After this, in outlier detection module, distributed algorithms is proposed on the preprocessed input data set for identifying outliers. These results will be evaluated for proposed algorithmic distributed approaches in the performance evaluation.

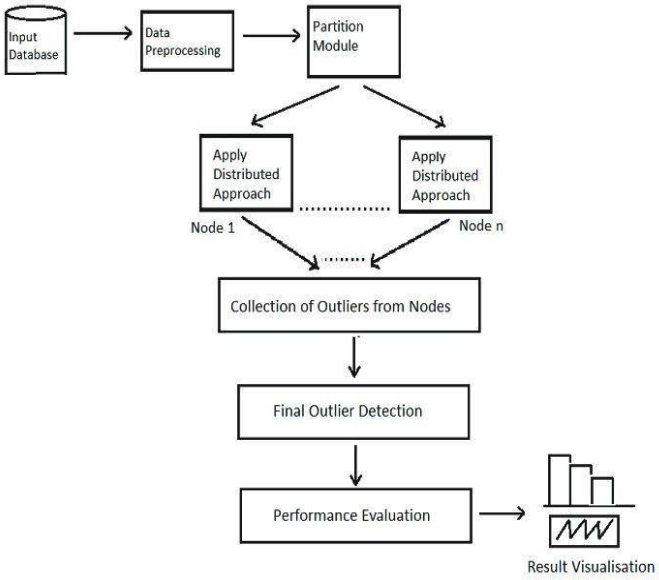
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017



1) Data collection and data preprocessing :

In data collection the initial input data for this system will be collected from standard dataset portal i.e. UCI data set repository. As proposed in system, the standard dataset will be used for this system includes Cover type, IPS datasets. Collected datasets may be available in their original, uncompressed form therefore; it is required to preprocess such data before forwarding for future steps. To preprocess large dataset contents, techniques available is data mining such as data integration, data transformation, data cleaning, etc. will be used and cleaned, required data will be generated.

2) Data partitioning:

In this module, as stated earlier in system execution plan, the preprocessed data is divided into number of clients from central supervisor node i.e. server as per the data request made by desired number of clients. This partitioned data will be then processed by individual clients to identify outliers based on applied algorithm strategy.

3) Outlier detection:

The technique proposed for identifying outliers will be applied initially at distributed clients and their results of detected outliers would be integrated on server machine at final stage computation of outliers. To try this, the outlier detection methods proposed are KNN Algorithm with ABOD and INFLO method.

The distributed procedure proposed with above approach established on anomaly detection techniques founded on nearest neighbor .In this procedure assumption is that normal data instances occur in dense neighborhoods, even as outliers occur a ways from their nearest neighbors. In this proposed work using principles of nearest neighbor founded anomaly detection techniques:(1) use the distance of an information example to its kth nearest neighbors to compute the outlier score (2) compute the relative density of each data instance to compute its outlier score.

The proposed algorithm consider the k-occurrences defined as dataset with finite set of n points and for a given point x in a dataset, denote the number of k-occurrences based on given similarity or distance measure as Nk(x),that the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

number of times x occurs among all other points in k nearest neighbor and points those frequently occurred as a hubs and points those occur infrequently as a antihub.Uses reverse nearest neighbors for instance , finding the instances to which query object is nearest. In this first read the each attribute in high dimensional dataset, then using angle based outlier detection technique compute the distance for every attribute using dataset Set distance and compare with distance from each instance and assign the outlier score. Based on that outlier score using reverse nearest neighbor determine that particular instance is an outlier or not.

4) Performance Evaluation and Result Visualization :

In this module, the outlier detected by above approach will be evaluated on the basis of set evaluation parameters for their performance evaluation. The performance evaluation will also provide details about implemented system performance metrics, constraints and directions for future scope. With the help of proper visualization of results, the system execution will be made more understandable and explorative for its evaluators.

V. CONCLUSION

This proposed KNN Algorithm with ABOD and INFLO Method with unsupervised learning using distributed approach aims at implement and comparing few of the unsupervised outlier detection methods and propose a way to improve them in terms of speed and accuracy, reducing the false positive error rate, reducing the false negative rate and improve the efficiency of density based outlier detection and comparison with the existing algorithms. The future implementation is in machine learning techniques such as supervised and semi-supervised methods.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput Surv, vol. 41, no. 3, p. 15, 2009.
[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” SIGMOD Rec, vol. 29, no. 2, pp. 93–104, 2000.
[3] 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 on Knowledge Discovery and Data Mining (PAKDD), pp. 813–822. 2009.
[4] W. Jin, A. K. H. Tung, J. Han, and W. Wang, “Ranking outliers using symmetric neighborhood relationship,” in Proc 10th Pacific-Asia Conf on Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 577–593, 2006.
[5] C. Lijun, L. Xiyin, Z. Tiejun, Z. Zhongping, and L. Aiyong, “A data stream outlier delection algorithm based on reverse k nearest neighbors,” in Proc 3rd Int Symposium on Computational Intelligence and Design (ISCID), pp. 236–239, 2010.
[6] Emmanuel Milller, Matthias Schiffer , Thomas Seidll,”Statistical Selection of Relevant Subspace Projections for Outlier Ranking”,IEEE, ICDE Conference,pp. 434 - 445 ,2011.
[7] Hans-Peter Kriegel Peer Kr¨oger Erich Schubert Arthur Zimek,” Interpreting and Unifying Outlier Scores”,SIAM International Conference on Data Mining (SDM), Mesa, pp. 13–24. AZ, 2011.
[8]Nattorn Buthong, Arthorn Luangsodsai, Krung Sinapiromsaran,” Outlier Detection Score Based on Ordered Distance Difference,” International Computer Science and Engineering Conference (ICSEC),pp. 157 – 162, 2013.
[9] H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in Proc 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD), pp. 444–452, 2008.