iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VII **Month of publication:** July 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A RNN Novel Approach for Unsupervised Distance-Based Outlier Detection

M. Siva Kumar[1], G. Prasadbabu M.E.[2] ,MISTE
[1]*M.Tech,* [2]*Associate professor , Siddarth Institute of Engineering and Technology*

I. INTRODUCTION

Detection of outliers in data defined as finding patterns in data that do not conform to normal behavior or data that do not conformed to expected behavior, such a data are called as outliers, anomalies, exceptions. Anomaly and Outlier have similar meaning. The analysts have strong interest in outliers because they may represent critical and actionable information in various domains, such as intrusion detection, fraud detection, and medical and health diagnosis. An Outlier is an observation in data instances which is different from the others in dataset. There are many reasons due to outliers arise like poor data quality, malfunctioning of equipment, ex credit card fraud. Data Labels associated with data instances shows whether that instance belongs to normal data or anomalous. Based on the availability of labels for data instance, the anomaly detection techniques operate in one of the three modes are 1)Supervised Anomaly Detection, techniques trained in supervised mode consider that the availability of labeled instances for normal as well as anomaly classes in a a training dataset. 2) Semi-supervised Anomaly Detection, techniques trained in supervised mode consider that the availability of labeled instances for normal, do not require labels for the anomaly class. 3) Unsupervised Anomaly Detection, techniques that operate in unsupervised mode do not require training data. There are various methods for outlier detection based on nearest neighbors, which consider that outliers appear far from their nearest neighbors. Such methods base on a distance or similarity measure to search the neighbors, with Euclidean distance. Many neighbor-based methods include defining the outlier score of a point as the distance to its *k*thnearest neighbor (k-NN method), some methods that determine the score of a point according to its relative density, since the distance to the kth nearest neighbor for a given data point can be viewed as an estimate of the inverse density around it.

Our motivation is based on the following factors:

It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in [10] the actual challenges posed by the "curse of dimensionality" differ from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space [9]. We will present further evidence which challenges this view, motivating the (re)examination of methods.

Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points [11], [12],1 but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data [14] warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method [11].

II. RELATED WORK

Author [2] assign an anomaly score known as Local Outlier Factor (LOF) to a given data instance. For any given data instance, the LOF score is equal to ratio of average local density of the *k* nearest neighbors of the instance and the local density of the data instance 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. In this to measure how much objects deviate from their scattered neighborhood. uses the relative distance from an object to its neighbors. The higher violation in degree of an object has, the mostly object is an outlier. In [4] proposed on a symmetric neighborhood relationship measure considers both neighbors and reverse neighbors of an object when estimating its density distribution .To avoid problem, when outliers are in the location where the density distributions in the neighborhood are significantly different. In [5] Author propose a data stream outlier detection algorithm SODRNN based on reverse nearest neighbor. Deal with the sliding window model, to detect anomalies outlier queries are performed in order in the current

window.

Improves efficiency by update of insertion or deletion only in one scan of the current window. In [6] propose a outlier ranking based on the objects deviation in a set of relevant subspace projections. It excludes irrelevant projections showing no clear difference between outliers and the residual objects and find objects deviating in multiple relevant subspaces, tackle the general challenges of detecting outliers hidden in subspaces of the data. In[7]Author propose a unification of outlier scores provided by various outlier models and a translation of the arbitrary "outlier factors" to values in the range [0,1] interpretable as values describing the probability of a data object of being an outlier. In [8] propose a new approach for parameter-free outlier detection algorithm to compute Ordered Distance Difference Outlier Factor. Formulate a new outlier score for each instance by considering the difference of ordered distances. Then, use this value to compute an outlier score.

III. EXISTING SYSTEM

A. The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Among these categories, unsupervised methods are more widely applied because the other categories require accurate and representative labels that are often prohibitively expensive to obtain.

B. Unsupervised methods include distance-based methods that mainly rely on a measure of distance or similarity in order to detect outliers. A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless, since distance measures concentrate, i.e., pair wise distances become indiscernible as dimensionality increases.

C. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good

IV. PROPOSED SYSTEM

It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in the actual challenges posed by the "curse of dimensionality" differs from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space. We will present further evidence which challenges this view, motivating the (re)examination of methods. Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data warrant their re examination for the outlier-detection task. In this light, we will revisit the ODIN method.

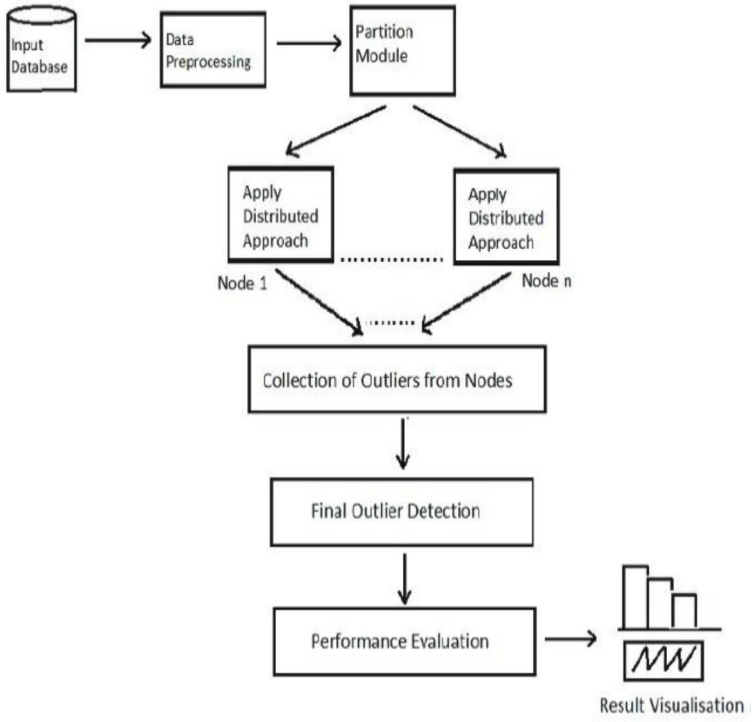


Fig1. System Architecture

A. 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.

B. 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.

C. 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 do this, the outlier
detection strategies proposed are KNN Algorithm with ABOD and INFLO Method.

The Distributed approach proposed with above Method based on anomaly detection techniques based on nearest neighbor .In this technique assumption is that normal data instances occur in dense neighborhoods, while outliers occur far from their nearest neighbors. In this proposed work using concepts of nearest neighbor based anomaly detection techniques:(1) use the distance of a data instance 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 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.

D. 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. EXPERIMENTAL SETUP AND EVALUATION

Our tests were performed using high dimensional dataset that is Cover Type dataset from UCI machine learning Repository which contains 54 number of attribute and number of instances are 581012.The experimental evaluation was performed on an Intel two core CPU at 2.53 GHz and 4 GB of RAM, having a windows as its operating system. The algorithm was fully implemented in Java to process data instances in high dimensional data.

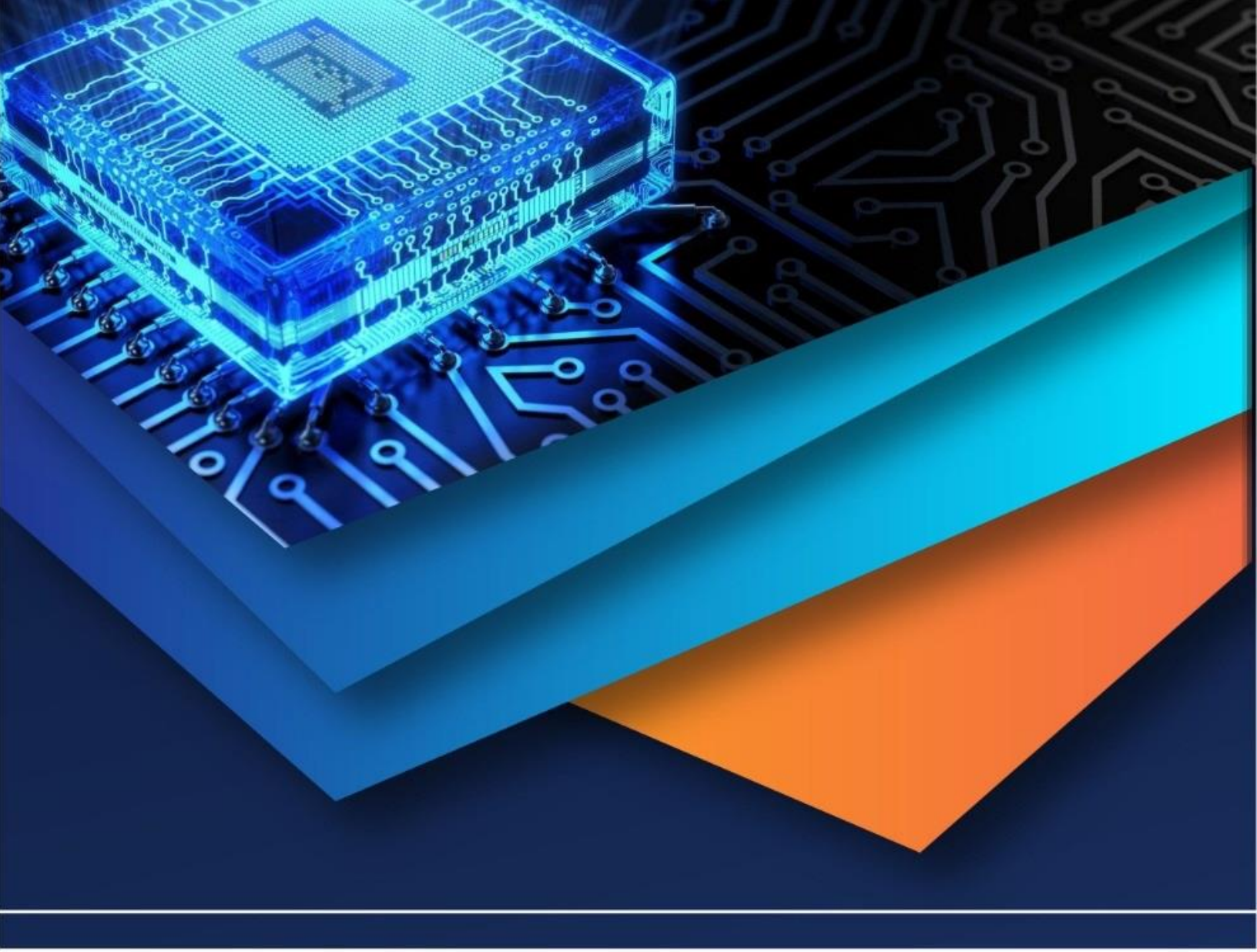
VI. CONCLUSION

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 Millller, 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY