IMPLEMENTATION OF DISTRIBUTED APPROACH FOR OUTLIER DETECTION USING NEAREST NEIGHBOR

Deepika Kankati[1], Ashish Ladda[2]
[1]M.Tech Student, BIES Engineering College, Warangal.
[2]Assistant Professor, BALAJI INSTITUTE OF TECHNOLOGICAL SCIENCES, Warangal.

Abstract: Outlier detection in high-dimensional information presents numerous challenges ensuing from the "curse of spatiality." A prevailing read is that distance concentration, i.e., the tendency of distances in high-dimensional information to become indiscernible, hinders the detection of outliers by creating distance-based strategies label all points as nearly equally smart outliers. During this paper, we offer proof supporting the opinion that such a read is just too easy, by demonstrating that distance-based strategies will turn out additional different outlier scores in high-dimensional settings. What is more, we have a tendency to show that top spatiality will have a special impact, by reexamining the notion of reverse nearest neighbors within the unattended outlier-detection context. Namely, it had been recently ascertained that the distribution of points' reverse-neighbor counts becomes skew in high dimensions, leading to the development called hub ness. We offer insight into however some points (anti-hubs) seem terribly occasionally in k-NN lists of alternative points, and justify the association between anti-hubs, outliers, and existing unattended outlier-detection strategies. By evaluating the classic k-NN technique, the angle-based technique designed for high-dimensional information, the density-based native outlier issue and influenced outlierness strategies, and anti hub-based strategies on numerous artificial and real-world information sets, we provide novel insight into the utility of reverse neighbor counts in unattended outlier detection.

Index Terms: Outlier detection, reverse nearest neighbors, high-dimensional data, distance concentration

I. INTRODUCTION

Outlier detection is studied wide within the survey as a result of would like of looking intrusion detection and anomaly detection in several applications. There square measure 3 main styles of outlier detection ways specifically, unsupervised, semi-supervised and supervised. These sorts square measure divided by labels of instances on that outlier detection is to be applied. would like availableness of correct labels of the instances for supervised and semi- supervised outlier detection. For outlier detection availableness of labels isn't much doable thus unsupervised technique is employed wide that doesn't would like label to the instances. most well-liked and effective technique for unsupervised outlier detection is distance based mostly outlier detection [1]. Distance based mostly outlier detection think about instances have tiny distance among them and outliers have massive distance from normal instances. V. Huhtamaki et al [2] expressed that because the dimensions of the info raises, distances flip useless to search out outliers as a result of every purpose looks as outlier. unsupervised outlier detection confronts some challenges in high-dimensionality. in spite of the common notion that each one points during a high-dimensional data-set appear to show outliers, Milos Radovanovic et al [20] showed that unsupervised ways will notice outliers below the idea that each one (or most) knowledge attributes square measure purposeful, i.e. not noisy. The relation between the high spatial property and outlier nature of the instances investigates by Milos Radovanovic et al [20]. K-nearest neighbor of {the purpose the purpose} P is K points whose distance to point P is a smaller amount than all different points. Reverse nearest neighbors (RNN) of purpose P is that the points that P is in their k nearest neighbor list. Some points square measure oftentimes comes in k-nearest neighbor list of different points and a few points square measure occasionally comes in k nearest neighbor list of another points square measure referred to as Anti-hubs. Density based mostly native Identifiers (LOF) [9] its variants square measure projected in literature. conjointly Angle-Based Outlier Detection is offered within the literature [10]. For outlier detection RNN thought is employed in literature [2] [4], however there's no theoretical proof that explores the relation between the outlier natures of the points and reverse nearest neighbors. Gustavo H. Orair et al[6] expressed that reverse nearest count is get affected because the spatial property of the info will increase, thus there's ought to investigate however outlier detection ways bases on RNN get littered with the spatial property of the info. Milos Radovanovic et al [20] discusses one. In high spatial property the issues in outlier detection and shows that however unsupervised ways may be used for outlier detection. 2. however Anti-hubs square measure relating to outlier nature of the purpose is investigates. 3. For outlier detection supported the relation anti-hubs and outlier 2 ways square measure projected for top and low dimensional knowledge for showing the outlierness of points, starting with the strategy Odin (Outlier Detection victimization in-degree Number). In existing system it takes massive computation value, time to calculate the reverse nearest neighbors of the all points. Use of ant hubs for outlier detection is of high process task. Computation complexness will increase with the info spatial property. For this there's scope to removal of impertinent options before application of Reverse Nearest Neighbor. thus to beat this downside, feature choice is applied on the info. during this step, all options square measure rank in step with their importance and needed options square measure selected for locating reverse nearest neighbors. to search out reverse nearest neighbor victimization geometrician distance and outlier score is calculated by victimization technique from existing system. in step with studies, if system doesn't realize the distribution of the info then geometrician distance is that the most suitable option. projected theme deals with curse of spatial property expeditiously. we tend to mentioned existing system, downside statement and projected theme with elaborated structure and algorithms.

II. PROBLEM STATEMENT:

A. Local outlier issue (LOF: In LOF, compare the native density of a instances with the densities of its neighborhood instances so assign anomaly score to given information instance. For Any information instance to be traditional not as an outlier, LOF score capable magnitude relation of average native density of k nearest neighbor of instance and native density of knowledge instance itself. to seek out native density for information instance, realize radius of tiny hyper sphere targeted at the information instance. The native density for instances is computed by dividing volume of k,i.e k nearest neighbor and volume of hyper sphere. during this assign a degree to every object to being AN outlier referred to as native outlier issue. Depends on the degree it determines however the thing is isolated with reference to close neighborhood. The instances lying in dense region area unit traditional instances, if their native density is comparable to their neighbors, the instances area unit outlier if there native density less than its nearest neighbor.LOF is a lot of reliable with top-n manner. Thence it's referred to as top-n LOF suggests that instances with highest LOF values think about as outliers.

B. native distance primarily {based} outlier issue(LDOF): native distance based outlier factor live the objects outlierness in scattered datasets . during this uses the relative location of AN object to its neighbors to see the thing deviation degree from its neighborhood instances. during this scattered neighborhood is taken into account. Higher deviation in degree information instance has, a lot of doubtless information instance as AN outlier. During this algorithmic program calculates the native distance based mostly outlier issue for every object so kind and ranks the n objects having highest LDOF price. the primary n objects with highest LDOF values area unit think about as AN outlier.

C. Influenced Outlierness (INFLO): This algorithmic program considers the circumstances once outliers area unit within the location wherever neighborhood density distributions area unit considerably completely different, maybe, within the case of objects near to a denser cluster from a thin cluster, this might offer wrong result. This algorithmic program considers the cruciate neighborhood relationship. during this considering influence house And once estimating its density distribution conjointly considers each neighbors and reverse neighbors of an object .Assign every object in a very information a influenced outlierness degree. the upper flow implies that the thing is AN outlier. D. Disadvantages: one. Threshold price is employed to differentiate outliers from traditional object and lower outlierness threshold price can end in high false negative rate for outlier detection . 2. downside arises once information instance is found between 2 clusters, the inhume distance between the thing of k nearest neighborhood will increase once the divisor price will increase results in high false positive rate. 3. must improve to reckon outlier detection speed. 4. must improve the potency of density based mostly outlier detection.

III. PROPOSED SOLUTION

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

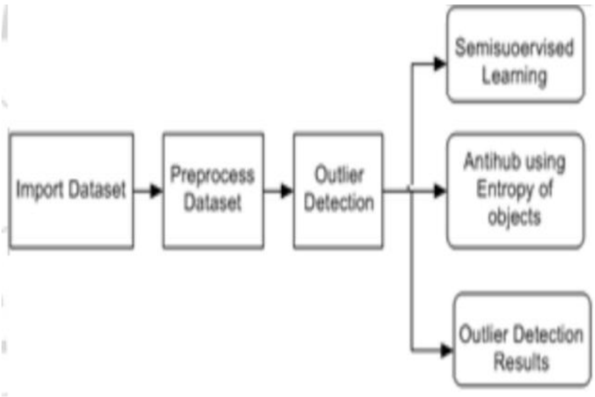

Fig 1: System architecture

1). Data collection and data preprocessing :

In information assortment the initial computer file for this method are going to be collected from commonplace dataset portal i.e. UCI information set repository. As planned in system, the quality dataset are going to be used for this method includes cowl kind, IPS datasets. Collected datasets could also be offered in their original, uncompressed kind therefore; it's needed to preprocess such information before forwarding for future steps. To preprocess massive dataset contents, techniques offered is data processing love information integration, information transformation, information cleansing, etc. are going to be used and cleansed, needed information are going to be generated.

2) information partitioning: during this module, as declared earlier in system execution set up, the preprocessed information is split into variety of purchasers from central supervisor node i.e. server as per the information request created by desired variety of purchasers. This divided

information are going to be then processed by individual purchasers to spot outliers supported applied formula strategy.

3) Outlier detection: The technique planned for distinctive outliers are going to be applied at first at distributed purchasers and their results of detected outliers would be integrated on server machine at last computation of outliers. To do this, the outlier detection methods planned square measure KNN formula with ABOD and INFLO methodology. The Distributed approach planned with higher than methodology supported anomaly detection techniques supported nearest neighbor .In this technique assumption is that standard information instances occur in dense neighborhoods, whereas outliers occur faraway from their nearest neighbors. during this planned work victimization ideas of nearest neighbor primarily based anomaly detection techniques:

(1) use the space of an information instance to its kth nearest neighbors to reckon the outlier score.

(2) reckon the density of every information instance to reckon its outlier score. The planned formula take into account the k-occurrences outlined as dataset with finite set of n purposes and for a given point x in a very dataset, denote the quantity of k-occurrences supported given similarity or distance live as Nk(x),that the quantity of times x happens among all different points in k nearest neighbor and points those often occurred as a hubs and points those occur sometimes as a Emmet hub. Uses reverse nearest neighbors as an instance , finding the instances to that question object is nearest. during this 1st scan the every attribute in high dimensional dataset, then victimization angle primarily based outlier detection technique reckon the space attribute victimization dataset Set distance and compare with distance from each instance and assign the outlier score. supported that outlier score victimization reverse nearest neighbor verify that exact instance is associate outlier or not. 4) Performance analysis and Result visual image: during this module, the outlier detected by higher than approach are going to be evaluated on the idea of set analysis parameters for his or her performance analysis. The performance analysis also will give details concerning enforced system performance metrics, constraints and directions for future scope. With the assistance of correct visual image of results, the system execution are going to be created a lot of intelligible and exploratory for its evaluators.

IV. CONCLUSION

In this paper, the survey is discussed with different ways in which problem of unsupervised outlier detection for high dimensional data has been formulated in literature and have attempted to provide an overview of huge literature on various techniques. Implementation of high dimensional data in most of the applications becomes an issue nowadays due to increase of dimensionality. Hubness is the recently known concept for handling the problems related with the increase of dimensionality and it is understood that it is an intrinsic property of the data where dimension is high. So the role of hubness has been examined in this paper. Also reviewed some of the recent advancements in unsupervised outlier detection for dealing with more complex high dimensional data with the usage of hubness. Outlier detection for high dimensional data is a fast growing emerging technique of today’s research and more new methods regarding this technique will emerge in the future.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu, “Outlier detection for high dimensional data” ACM SIGMOD RECORD February 2002.

[2] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey ACM Comput Surv, vol. 41, no. 3, p. 15, 2009.

[3] Zhang, Ji. "Advancements of outlier detection: A survey." ICST Transactions on Scalable Information Systems 13.1 (2013): 1-26.

[4] Milo˘s Radovanovi´c, Alexandros Nanopoulos, and Mirjana Ivanovi´ “Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection” IEEE Transactions On Knowledge And Data Engineering, October 2014.

[5] 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 and Data Mining (KDD), 2008, pp. 444–452.

[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, n4o. 2, pp. 309–336, 2011.

[7] H.P. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in Proc 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.

[8] 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.

[9] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection for discrete sequences: A survey,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 5, pp. 823–839, May 2012.

[10] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, “Svdd-based outlier detection on uncertain data,” Knowl. Inform. Syst., vol. 34, no. 3, pp. 597–618, 2013.

[11] Hans-Peter Kriegel,Matthias Schubert and Arthur Zimek, ”Angle-Based Outlier De-tection in Highdimensional Data,” 2008.

[12] D. Francois, V. Wertz, and M. Verleysen, ”The concentration offractional distances,”IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7,pp. 873-886, Jul. 2007.

[13] A. Nanopoulos, Y. Theodoridis, and Y. Manolopoulos, ”C2P:Clustering based on closest pairs,” in Proc 27th Int. Conf. VeryLarge Data Bases, 2001, pp. 331-340

Kankati deepika Currently doing M.Tech in computer science and Engineering at BIES Engineering College, Warangal, India. Research interests include Networks, Mobile Computing etc.,

Ashish Ladda is 4+ years experienced Assistant Professor in the department Computer Science & Engineering, BALAJI INSTITUTE OF TECHNOLOGICAL SCIENCES-NARSAMPET, Warangal, India and his Research area includes Data Mining , Network Security etc.,