Outlier Detection Using Anti-hubs

Miss.Gavale Swati S.[1] Prof. Kahate Sandip[2]
[1]M. E Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Sharadchandra Pawar College of Engineering, Dumbarwadi Otur, Pune, India

Abstract— The distance base outliers detection method fails to increase the dimensionality of the data. This problem occurs due to irrelevant and redundant feature because the distance between two points is less. Reverse nearest neighbors of point P is the points for which P is in their k nearest neighbor list. Antihubs are some points are frequently comes in k-nearest neighbor list of another points and some points are infrequently comes in k nearest neighbor list of different points. Latest proposes are antihub base unsupervised outlier detection method, but these propose are suffering from high computational cost of finding outlier. This is depends on data who having better dimensionality, high computation cost, time requirement to find high antihubs. To avoid this there is need to remove out irrelevant and redundant feature of high dimensionality data. It increases the efficiency by removing the redundant feature. Using feature selection method redundant feature are removed.

Key words: Nearest Neighbor, Outlier Detection, Reverse Nearest Neighbors

I. INTRODUCTION

There is need of finding intrusion and outlier detection methods like unsupervised, semi supervised and supervised by studying outlier detection. These types are divided into labels of instances on which outlier detection is to be applied.There is need to availability of correct labels of instance for supervised and semi supervised outlier detection. Distance based outlier detection most popular and effective method for unsupervised outlier detection. In the distance base outlier detection normal distances have small distance among them and outliers have large distance between them.As the dimensionality increases the distance fails to find outliers.

Unsupervised methods can detect outliers under the assumption that all data attributes are purposeful, i.e. not noisy. The relation between the high dimensionality and outlier nature of the instances investigates. Some points are frequently comes in k-nearest neighbor list of other points and some points are infrequently comes in k nearest neighbor list of some other points are called as Anti-hubs.

For outlier detection RNN concept is used but there is no theoretical proof which explores the relation between the outlier natures of the points and reverses nearest neighbors. S. Ramaswamy et al [6] stated that reverse nearest count is get affected as the dimensionality of the data increases, so there is need to investigate how outlier detection methods bases on RNN get affected by the dimensionality of the data.

1) In high dimensionalitythe problems in outlier detection and shows that how unsupervised methods can be used for outlier detection.
2) How Anti-hubs are related to outlier nature of the point is investigates.
3) For outlier detection based on the relation anti-hubs and outlier two methods are proposed for high and low dimensional data for showing the outlier-ness of points, beginning with the method ODIN (Outlier Detection using in-degree Number).

Existing system, it takes large computation cost, time to calculate the reverse nearest neighbors of the all points. Use of antihubs for outlier detection is of high computational task. Computational complexity depends on data dimensionality as dimensionality of data increases the complexity of computation increases. Because of this nearest neighbor introduced to remove irrelevant and redundant data. To avoid this feature selection is introduced. All the features are arrange rank wise and required features are taken for reversed nearest neighbor. Outlier score is calculated by using reversed nearest neighbor. According to studies, if system does not know about the distribution of the data then euclidean distance is the best choice.

II. LITERATURE SURVEY

Edwin M. Knorr et al [1] described to find outlier in heavy multidimensional dataset. Existing system is used to find outliers which can only deals with multi dimension attribute of dataset here outlier detection could be done efficiently for heavy dataset and k dimensional dataset along with large value of k and outlier detection of useful with clear meaning and useful knowledge gain task. For finding outlier proposed and analyze some algorithm.

Amit Singh et al[2] stated that reverse nearest queries is widely used in such an applications decision support system, data streaming documents profile base marketing bioinformatics. To solve this problem is high dimensional data. In this paper proposed solution is used for to reverse nearest neighbor queries in high dimensional dataset using k nearest neighbor and reverse nearest neighbor. The problem of finding RNN in high dimensions not covered in the past. They discussed the challenges and perfected some important observations related to high dimensional RNNs. Then proposed an approximate solution to answer RNN queries in high dimensions.

Ke Zhang et al[3] described outlier detection has some issues with some dataset is a large challenge in this world in KDD application. Existing system of outlier detection is not effective on scattered dataset due to which constant pattern and parameter setting problem. Here paper introduced a local distance based outlier factor (LDOF) to calculate outliers of object in scattered dataset.

Mahbod Tavallaee et al[4] stated that during anomaly detection researcher has many problem occur in the detection of novel attacks and KDDcup 99 there is weakness of signature based IDSs .Those dataset are very useful and widely used in analysis. To solve this issues this paper proposed a new dataset NSL-KDD which consist of required records which are redundant records are removed and not

goes from any attacks. The analysis shows performance of evaluated systems and their results.

Markus M. Breunig et al[5] described ,in many KDD applications finding outliers is more interesting than finding common patterns. Outlier is a binary value (0, 1) in outlier detection. In this paper it is more useful to give each object a degree as outlier .This degree is called local outlier factor (LOF) of an object. This degree is depends on object which is placed at neighborhood. Then performance evaluation of our algorithm confirms to show that our approach of finding local outliers can be practical.

Sridhar Ramswamy et al[6] proposed various attack formation using distance base outliers .This is calculated using distance of appoint from kth nearest neighbor. Using rank basis to its distance outlier can be calculated .For this experimental study real life and synthetic dataset are used. Analyze partition based algorithm for mining outliers.

Hans-Peter Kriegel et al[7] use different technique for finding outlier detection in different group of dataset is the main task. Existing approach are to find outliers using distance in full dimensional datasets, in high dimensional datasets this approaches degrades due to curse of dimensionality. In this paper we proposed angle based outlier detection (ABOD) angle based outlier detection and access the angle between points of a vector to another point. In this way effect of curse of dimensionality is recover as compare to distance approaches. Also verified the performance of ABOD is better than another and its useful for high dimensional data.

Charu C. Aggarwal et al[8] described, in past there is some issues like curse of high dimensionality in nearest neighbor and indexing is high dimensional data become sparse and indexing fails from efficiency and therefore the dataset records are not useful or bad quality . They studied dimensionality curse as the point of view the distance metrics which is used to measure similarity between objects. They show that the fractional distance metrics provide more useful results. Results can improve the effect of standard clustering and knn algorithms.

Edwin M. Knorr et al [9] stated that existing system are outlier focuses on the identification aspects.None provide any intentional knowledge of the outliers.Evaluate the validity of the identify outliers and improve one's understanding of data. In this ,they showed that what kind of intentional knowledge provide and how to optimize computation of such knowledge for first issue proposed strongest and week outlier and for second proposed naive and semi naive algorithm. For the experimental analysis result shows the significance reduction in input and output and significant speedup in runtime.

Stephen D. Bay et al[10] declared outliers by their distance to neighboring examples are a popular approach to finding unusual examples in a dataset. Stephen D. Bay et al use a simple nested loop algorithm that in the worst case is quadratic can give near linear time performance when the data is in random order and a simple pruning rule is used. Stephen D. Bay et al test algorithm on real high-dimensional data sets with millionsof examples and show that the near linear scaling holds over several orders of magnitude. Average case analysis suggests that much of the efficiency is because the time to process non-outliers, which are the majority of examples, does not depend on the size of the data set.

Milos Radovanovic et al [11] discoursed issues in outlier detection in the case of eminent data dimensionality and showed the way outlier detection in high dimensional data can be made using unsupervised methods describe. It also enquires how Anti-hubs are associated to the point's outlier nature.

III. PROBLEM DEFINITION

With the curse of dimensionality, computation complexity increases. Existing method for outlier detection using reverse nearest neighbour suffers from high computation requirement for RNN. High computation results into reduced time efficiency and high memory requirement. There is need to overcome this issue of performance and computation in RNN process.

IV. EXISTING SYSTEM

From set of instances existing system consist of the process of finding irregular instances and it aims at make the use of outlier detection in finding intrusion detection and outlier detection in many applications and real life sources. Existing system discussed the problem in outlier detection in high dimensionality and shows that how unsupervised methods can be used for outlier detection in high dimensional data. It also investigates how anti-hubs are related to outlier nature of the point and based on the relation anti-hubs and outlier, there are two ways of using k-occurrence information are proposed for outlier detection for high and low dimensional data for showing the outlierness of points, beginning with the method ODIN (Outlier Detection using in-degree Number).

Limitation of existing System

- To calculate the reverse nearest neighbors of the all points it takes high computation cost, time in existing system.
- For outlier detection use of antihubs is of high computational task
- Computation complexity depends on the data dimensionality.

V. PROPOSED SYSTEM

To remove the drawback of the existing system, proposed system architecture contains three main steps:-

1) Feature Selection
2) Find Reverse nearest neighbor
3) Find outlier score of each instance

A. Feature Selection:

To handle the effect of curse of dimensionality proposed system is designed. Existing system required high computation cost, time to calculate the reverse nearest neighbors of the all points. Feature selection method is applied on the datato recover this problem. All features are rank according to their importance and required features are selected for finding reverse nearest neighbors.

B. Find Reverse Nearest Neighbor:

Data of selected features will be considered for searching the reverse nearest neighbor. To evaluate the reverse nearest neighbor, first k-nearest neighbors of each point is evaluated.

From the k-nearest neighbor list of each point, reverse nearest neighbor list of each point is evaluated.

C. Outlier Score of Each Point:

Previous methods than existing system considered k-occurrence of the point as an outlier score. Less k-occurrence indicates more outlier score of the point. Proposed system will follow existing system to calculate the outlier score of the point. Sum of k-occurrence score of k-nearest neighbors of the point P is outlier score of the point P.

VI. CONCLUSION

In many applications there is need of finding intrusion or outlier detection. Exiting conclude that reverse nearest neighbor outlier detection using anti-hub. But using anti hub for outlier detection is of high computational task. Computational complexity increases with the data dimensionality to avoid this removal of irrelevant features before application of reverse nearest neighbor is introduced. From actual results it is clear that proposed system maintains the accuracy and also reduces the time and memory requirement for outlier detection.

REFERENCES

[1] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237–253, 2000.
[2] A. Singh, H. Ferhatosmano_glu, and A. ¸Saman Tosun, "High dimensional reverse nearest neighbor queries," in Proc 12th ACMConf. Inform. Knowl. Manage., pp. 91–98, 2003.
[3] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc 13thPacific-Asia Conf. Knowl. Discovery Data Mining, pp. 813–822, 2009.
[4] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. 2nd IEEE Symp.Comput. Intell. Secur. Defense Appl., pp. 1–6, 2009.
[5] 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.
[6] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms formining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2,pp. 427–438, 2000
[7] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc 14th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, pp. 444–452,2008.
[8] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," inProc. 8th Int. Conf. Database Theory, pp. 420–434, 2001.
[9] Edwin Knorr and Raymond Ng. "Finding intensional knowledge of distance-based outliers". In Proc. of the VLDB Conference, ages 211–222, Edinburgh, UK, September 1999.
[10] Edwin Knorr and Raymond Ng. "Algorithms for mining distance-based outliers in large datasets". In Proc. of the VLDB Conference, pages 392–403, New York, USA, September 1998
[11] Milos Radovanovic, Alexandros Nanopoulos, and MirjanaIvanov,"Reverse nearest Neighbors in Unsupervised Distance-Based Outlier Detection," 2015.

