ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 6, June 2016

Comparing Accuracy,Time and Memory based on Unsupervised Outlier Detection using Anti-hubs

Miss.Gavale Swati S., Prof. Kahate Sandip

Abstract— **Outlier detection method is one of the best methods for outlier detection but as the number of dimensions increases this method fails to detect outliers. Unwanted feature are the problem for this process; the distance of nearest neighbor is less than other point. Those points which are frequently comes k-nearest neighbor list of another point are called hubs and infrequently comes k-nearest neighbor list of another point are called anti-hubs. There are many researches of antihub based unsupervised outlier detection but there is one issue is occurring that high computation cost for finding antihubs. If the data that having better dimensionality, high computation cost, high computation complexity, time requirement to find antihubs. To remove the unwanted feature and high dimensionality data to make an efficient system results. Feature selection is the process proposed to remove unwanted feature and make a system more efficient. We used R2l and U2r are two different dataset for the compare the performance of our system. Two different dataset gives different results for the analysis.**

Index Terms— **Data outlier detection, nearest neighbor, high dimensional data, reverse nearest neighbors.**

• INTRODUCTION

There are three main types of outlier detection methods namely, unsupervised, semi-supervised and supervised. There is need to find outlier in many application for that we have to study outlier detection analysis. There is need of availability of correct labels of the instances for supervised and semi- supervised outlier detection. Currently unsupervised technique is used widely which does not need label to the instances for outlier detection. Currently the best efficient method for outlier detection is unsupervised distance base outlier detection method. The normal instances have small amount of distances among them and outliers have large amount of distances among them in distance base outlier detection. As the increase in dimensions of data distances are not useful to find outliers causes the every point become an outliers. Therefore high dimensionality of data is the biggest problem or challenge for unsupervised distance outlier detection. Milos Radovanovic et al can show that unsupervised distance base outlier detection system can handle high dimensional data, it can detect outlier under specific condition i.e. data should be useful and attributes are meaningful means data should not be noisy then it can be success to handle high dimensional data. K-nearest neighbor (KNN) of the point P is K points whose distance to point P is less than all other points. Reverse nearest neighbors (RNN) of Point P is the points for which P is in their k nearest neighbor list. Some points are frequently comes in k-nearest neighbor list of other points are called hubs 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 in literature, but there is no theoretical proof which explores the relation between the outlier natures of the points and reverses nearest neighbors. The 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. Milos Radovanovic et al [17] discusses the problems in outlier detection for high dimensionality and shows that how unsupervised methods can be used for outlier detection. How anti-hubs are related to outlier nature of the point is investigates. For outlier detection based on the relation anti-hubs and outlier two methods are proposed for high and low dimensional data for showing the outlierness of points, beginning with the method ODIN (Outlier Detection using in-degree Number). Existing system, it has high computation cost, more time to calculate the reverse nearest neighbors of the all points. Use of antihubs for outlier detection is of high computational task and high complexity. Computation complexity increases with the data dimensionality increases. For this there is scope to removal of useless attribute and unwanted features before application of Reverse Nearest Neighbor. So to overcome this issue, feature selection is providing to the data. All features are arranging according to their importance and required features are selected for finding reverse nearest neighbors (RNN). To find reverse nearest neighbor using euclidean distance and outlier score is calculated by using technique from existing system by supervised method. According to this, if system does not know about the distribution of the data then euclidean distance is the best option. Proposed scheme deals with curse of dimensionality efficiently. We will study literature survey and implementation in detail.

Manuscript received June, 2016.
Miss. Gavale Swati S.*, Computer Engineering Department, Savitribai Phule Pune University, Pune, India,*
Prof.Kahate Sandip, *Computer Engineering Department, Savitribai Phule Pune University, Pune, India,*

- LITERATURE SURVEY

The problems arises due to increase in dimensionality of the data investigated by M. E. Houle et al [1]. Poor discrimination was caused by presence of the redundancy of attributes, presence of the irrelevant features and concentration. These issues reduce the usability of the similarity and distance measures. They evaluated that secondary measures like shared-neighbor were still useful in such condition.

V. Hautamaki et al [2] used reverse nearest neighbor count to score outlier nature of the point. User defined threshold was used to take decision about outlier nature of the point. Method proposed by V. Hautamaki et al named as Outlier Detection using in degree Number (ODIN). If score is less than threshold then the point is said to be an outlier otherwise it is normal point. The link between the reverse nearest neighbor count and outlier nature of the point investigate by V. Hautamaki et al.

Edwin M. Knorr et al[3]describes 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.

Reverse nearest queries is widely used in such an applications decision support system, data streaming documents, profile base marketing bio informatics. To solve this problem is high dimensional data. Amit Singh et al [4] 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 they proposed an approximate solution to answer RNN queries in high dimensions.

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. Ke Zhang et al [5] introduced a local distance based outlier factor (LDOF) to calculate outliers of object in scattered dataset.

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 Mahbod Tavallaee et al [4] 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.

Many current systems uses data mining based methods or the methods that are based on signature which depends on labeled data i.e. supervised training data for the purpose of intrusion detection described by E. Eskin et al [6]. As such systems can only detects the intrusions based on previously identified intrusions, there was a risk of attack until new type of attack has been manually revised. To train the model which will detect the attacks, anomaly detection algorithms based on supervised learning needs purely normal data, might contain some intrusions. Algorithm may not be able to identify and predict the future instances of such intrusions as it was considered as normal. To address the problem, they proposed a geometric model for anomaly detection founded on unsupervised method. For the detection of the outlier, E. Eskin et al proposed three algorithms. First algorithm was based on cluster based technique, second was based on k-nearest neighbor and third was based on support vector machine based algorithm.

The context of outlier detection, in this approach assigned the each object with the level of being an outlier and this assigned level i.e. degree of the object was called as Local Outlier Factor (LOF) explored by H.P.Kriegel et al [7]. In this approach, they used density based clustering for finding outliers in multidimensional datasets. Local Outlier factor means instead of considering an outlier as a dual dimension, assign object a level to which it was kept apart from the around neighbors.

Hans-Peter Kriegel et al [8] describe distance based approaches. This distance based approaches degrades in performance due to high dimensionality of the data. As they rely on the distances curse of dimensionality arises the performance issues. Identification of Density Based Local Outliers (LOF) degrades in terms of accuracy. To overcome the issues they introduced novel approach known as Angle-Based Outlier Detection (ABOD). Instead of considering only the points in the vector space also considered the directions of the distance vectors and showed that an outlier can be anticipated by equating the angle amongst the two distance vectors to other points. Proposed approach is not depends upon any method of selecting parameter which increased the process result achieving quality. With ABOD approach, two variants are proposed, named as Fast ABOD worthy for big data sets with low dimensions and LB-ABOD suitable for high dimensional data.

Numeral data analysis tools and nearest neighbor search mostly based on the use of euclidean distance describe by D. Franc et al [11]. In case of broad dimensionality, though all distances amongst different couple of date elements appears similar; the euclidean distance appear to concentrate. Therefore the distance's relevancy has been doubted in the past, and fractional norms were brought to overcome this problem. They suggested the use of alternative distances to agitate the concentration.

Outliers do not follow the distributed data, and may leads bad results with respect to statistical analysis. Many outlier detection tools are based on the assumption that the data is identically and independently distributed. Hancong Liu et al [12] proposed an outlier-resistant data filter-cleaner. The proposed data filter-cleaner includes an on-line outlier-resistant estimate of the process model and combines it with a modified Kalman filter to detect and "clean" outliers. The advantage over existing methods is that the proposed method has the following features: (a) a priori knowledge of the process model is not required; (b) it is applicable to auto correlated data; (c) it can be implemented on-line; and (d) it tries to only clean (i.e., detects and replaces) outliers and preserves all other information in the data.

Milos Radovanovi et al [13] proposed various methods of the curse of dimensionality are known to present serious challenges to various machine-learning methods and tasks. They discussed a new method of the dimensionality curse,

referred to as hubness that affects the distribution of k-occurrences: the number of times a point appears among the k nearest neighbors of other points in a data set. Through theoretical and empirical analysis involving synthetic and real data sets they show that under commonly used assumptions this distribution becomes considerably skewed as dimensionality increases, causing the emergence of hubs, that is, points with very high k-occurrences which effectively represent "popular" nearest neighbors. They examine the origins of this phenomenon, showing that it is an inherent property of data distributions in high-dimensional vector space, discuss its interaction with dimensionality reduction, and explore its influence on a wide range of machine-learning tasks directly or indirectly based on measuring distances, belonging to supervised, semi-supervised, and unsupervised learning families.

Outlier scores provided by different outlier models differ widely in their meaning, range and contrast between different outlier models and hence that are not easily comparable or interpretable. Hans-Peter Kriegelet al[14] proposed 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. As an application, Hans-Peter Kriegel et al show that this unification facilitates enhanced ensembles for outlier detection.

Outlier mining is very useful task to distinguish exceptional outliers from regular objects. Outlier mining in the full data space, there are well established methods which are successful in measuring the degree of deviation for outlier ranking. In recent applications traditional outlier mining approaches miss outliers as they are hidden in subspace projections. Outlier ranking approaches measuring deviation on all available attributes miss outliers deviating from their local neighborhood only in subsets of the attributes. Emmanuel Muller et al [15] proposed a novel outlier ranking based on the objects deviation in a selected set of relevant subspace projections. In thorough experiments on real and synthetic data they show that our approach outperforms competing outlier ranking approaches by detecting outliers in arbitrary subspace projections.

Hermine N. Akouemo et al [16] proposed the combination of two statistical techniques for the detection and imputation of outliers in time series data. An autoregressive integrated moving average with exogenous inputs (ARIMAX) model is used to extract the characteristics of the time series and to find the residuals. The outliers are detected by performing hypothesis testing on the residuals and the anomalous data are imputed using another ARIMAX model. They tests the algorithm using both synthetic and real data sets and present the analysis and comments on those results.

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 described by Milos Radovanovic et al[17]. It also enquires how Anti-hubs are associated to the point's outlier nature.

- IMPLEMENTATION

Implementation of proposed system is completed in different modules. In this very first stage is preprocessing of dataset. Here we used KDD dataset for the analysis but dataset contains all types of attributes value from which we required some attributes values from given dataset. In which we preprocess dataset, here first we take required attribute values of dataset but it is not normalized again preprocessing it converts to normalize. In preprocessing first we select column which having string values then get the distinct value from the given column after that the unique string value are replaced by unique id value and complete the process. Then store the given dataset file into target folder. Here we got the final dataset for the analysis.

The feature Selection phase is proposed to reduce time and memory and system made more efficient. Here we calculate required value from the formula. This is used to calculate MI value,

$$MI(A, B) = \sum_a \sum_b P_{AB}(a, b) \log \frac{P_{AB}(a,b)}{P_A(a)P_B(b)} \quad \ldots\ldots\ldots\ldots (1)$$

After calculating all required values we put values in the formula and got the MI value. Then we calculate threshold by taking average value of all the MI. To remove the negative or unwanted value average value is selected. After which we take the entire feature which having higher MI value than the threshold. Then selected features are store into a new file for the further processing.

After calculating selected feature reverse nearest neighbor of each point is evaluated. The distance between each point calculated. Before calculating reverse nearest of each point the K-nearest neighbor of each point is calculated. Euclidian measure is used to calculate distance between two instances. From the k-nearest neighbor list of each point, reverse nearest neighbor list of each point is calculated. Then List is sorted by ascending order for easy to understand.

The first antihub algorithm implemented in which all the reverse nearest neighbor is input to the antihub. Here outlier sore is calculated by using (1/size of the entire dataset). Then the sum of all outlier sore occurrences is calculated. The outlier score threshold is use to find out the outlier feature. Here actually the outliers features are having 1 at the end of the line and 0 having are normal instances. From which the count is calculated. Count/total features size is use to calculated the accuracy of the implementation. Here we also evaluated the values of total true positive, true negative, false positive, false negative for the analysis.

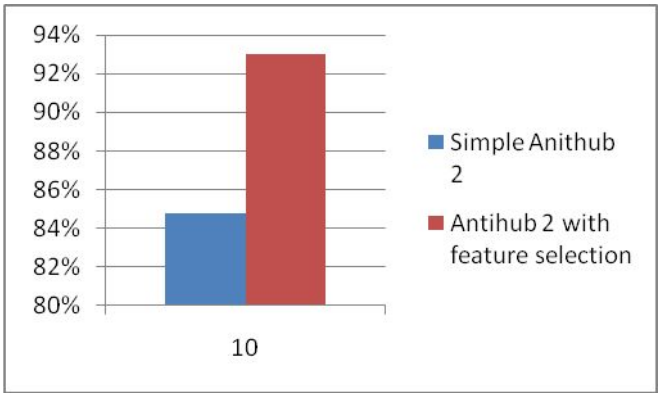
Outlier detection implementation using feature selection gives the better performance than the normal outlier detection implementation. Feature selection method require less time, less memory than the normal outlier detection process. Accuracy using feature selection process is more than the normal outlier detection process. Finally the outlier detection process by using feature selection is more efficient and gives better performance than normal outlier detection process.

- COMPARISION WITH OTHER DATA SET BASED ON ACCURACY, TIME AND MEMORY

Two different dataset are gives to our system for the comparison of accuracy, time and memory based on unsupervised outlier detection using antihubs. R2l and U2r two different dataset are used for outlier detection system. In which r2l having 2000 records (instances) and 42 attributes and u2r having 2000 records (instances) and 42 attributes.

TABLE 1.ACCURACY COMPARISON WITH K VARIATION

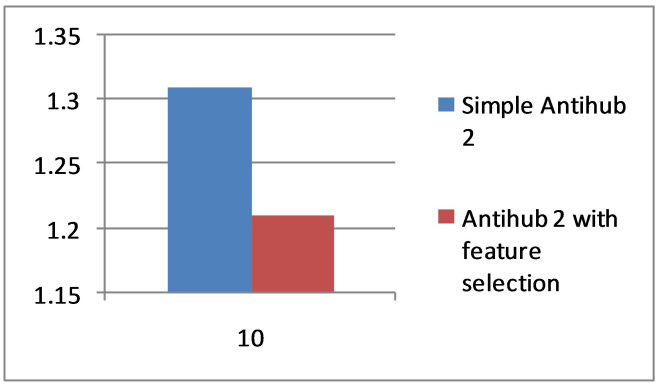
K	Simple Anithub 2	Antihub 2 with feature selection
10	84.79 %	93.40%



- Accuracy comparison with k variation

TABLE 2.TIME COMPARISON WITH K VARIATION

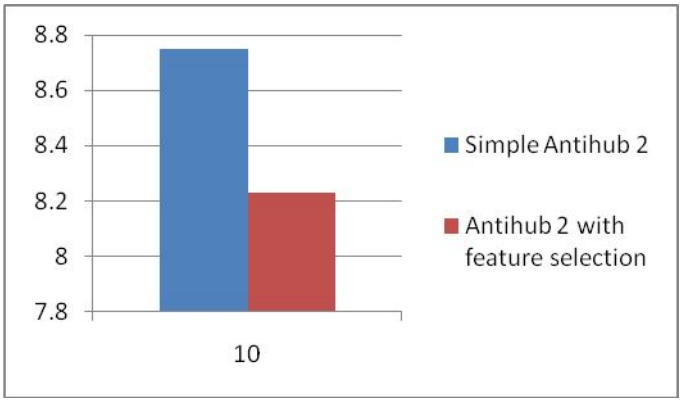
K	Simple Anithub 2 time in minute.	Antihub 2 with feature selection time in minute.
10	1.31	1.21



- Time comparison with k variation

TABLE 3.MEMORY COMPARISON WITH K VARIATION

K	Simple Anithub 2 memory in MB	Antihub 2 with feature selection memory in MB
10	8.75	8.23



- Memory comparison with k variation

For dataset U2R Results are shown below:-

TABLE 1.ACCURACY COMPARISON WITH K VARIATION

K	Simple Anithub 2	Antihub 2 with feature selection
10	95.79 %	99.40%

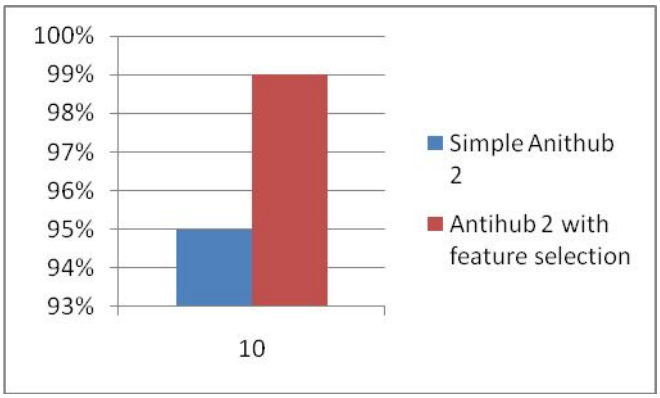


Fig. 1. Accuracy comparison with k variation

TABLE2.TIME COMPARISON WITH K VARIATION

K	Simple Anithub 2 time in minute.	Antihub 2 with feature selection time in minute.
10	3.11	3.05

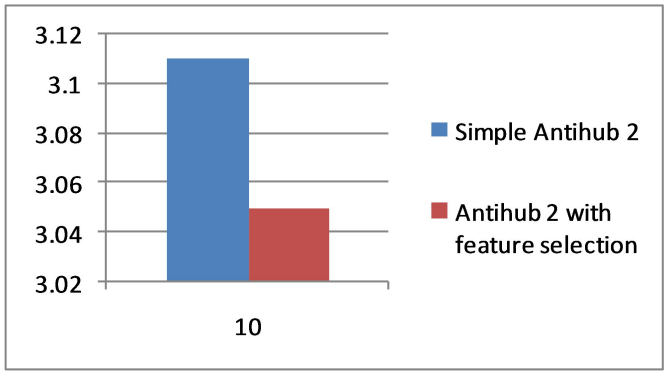


FIG. 2. TIME COMPARISON WITH K VARIATION

Table 3.Memory comparison with k variation

K	Simple Anithub 2 memory in MB	Antihub 2 with feature selection memory in MB
10	21.26	19.90

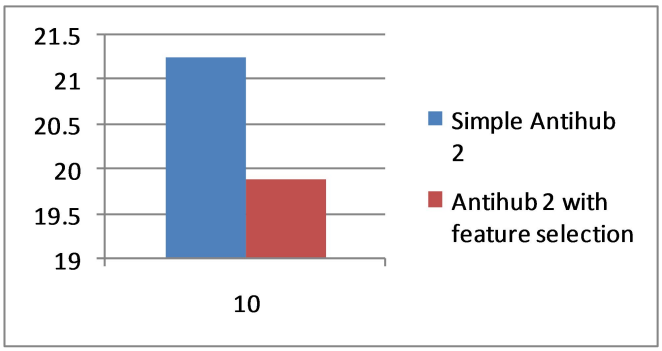


Fig.3. Time comparison with k variation

• CONCLUSION

Exiting method investigated reverse nearest neighbor outlier detection using anti-hubs. But using anti hubs for outlier detection is complex and time consuming computational task. Because of large number of computations, computational complexity increases with the data dimensionality. To avoid this unwanted features are discarded before application of reverse nearest neighbor. From experimental results, it is clear that proposed system increases the accuracy with decrease in time and memory requirement for outlier detection.

• FUTURE SCOPE

For achieving the better experimental results in future, we upgrade the proposed system such that it can handle large dimensional data and large computation complexity and also make more efficient system for finding the intrusion and outlier.

REFERENCES

[1] M. E. Houle, H.P.Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?," inProc 22nd Int. Conf. Sci. Statist.DatabaseManage., 2010, pp. 482-500.

[2] V. Hautamaki, I. Karkkainen, and P. Franti,"Outlier detection using k-nearest neigh-bour graph," in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430-433.

[3] 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.

[4] A. Singh, H. Ferhatosmano_glu, and A. ¸Saman Tosun, "High dimensional reverse nearest neighbor queries," in Proc 12th ACMConf. Inform. Knowl. Manage., 2003, pp. 91–98.

[5] 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, 2009, pp. 813–822.

[6] 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., 2009, pp. 1–6.

[7] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?,"in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217-235.

[8] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data,"in Proc. Conf. Appl. Data Mining Comput. Security, 2002, pp. 78-100.

[9] 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.

[10] Hans-Peter Kriegel,Matthias Schubert and Arthur Zimek, "Angle-Based Outlier De-tection in High-dimensional Data," 2008.

[11] 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.

[12] P. J. Rousseeuw and A. M. Leroy, "Robust Regression and Outlier Detection" Hoboken, NJ, USA: Wiley, 1987.

[13] M. Radovanovi_c, A. Nanopoulos, and M. Ivanovi_c, "Hubs in space: Popular nearest neighbors in high-dimensional data," J. Mach. Learn. Res., vol. 11, pp. 2487–2531, 2010.

[14] H.-P. Kriegel, P. Kr€oger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in Proc 11th SIAM Int. Conf. Data Mining, 2011, pp. 13–24.

[15] E. M€uller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in Proc. 27th IEEE Int. Conf. Data Eng., 2011, pp. 434–445.

[16] Hermine N. Akouemo and Richard J. Povinelli "Time series outlier detection and imputation" Milwaukee, Wisconsin 53233,july 2014

[17] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanov , "Reverse nearest Neighbors in Unsupervised Distance-Based Outlier Detection," 2015.