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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Reverse Nearest Neighbors in Unproven Reserve Outlier Discovery Based on Distance

Peddinti Rama Mohana Rao, Dr.N.Gopala Krishna Murthy

Dept. of Information Technology, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India

ABSTRACT: As the dimensionality of the data increases due to this all point becomes good outlier the distance based outlier detection methods fails. Reason of these issues is irrelevant and redundant features; nearest neighbor 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 another points referred as hubs and some points are infrequently comes in k nearest neighbor list of different points are called as Anti-hubs. Recent research proposes anti-hub based unsupervised outlier detection methods but these propose are suffered from computation cost of finding anti-hubs. In the case of data which has outstanding dimensionality, computation cost and time requirement to find anti-hubs is high. There is need to remove the redundant features if high dimensional data contains redundant attributes. Reduce the computation cost and time requirement by removal of redundant features to find anti-hubs for outlier detection. For extending anti-hub based outlier
Detection method for high dimensional data applies feature selection.
.

KEYWORDS: High-Dimensional, Data Outlier Detection, Reverse nearest Neighbors.

I. INTRODUCTION

Outlier detection is the method which identifying Patterns that do not conform to established standard behavior. Hawkins defines “the outlier as observation that deviates to large extent from the other observation which means that the pattern is generated by the different mechanism” Outlier detection is the process of finding knowledge from large and multidimensional databases to learn the unexpected pattern

and behavior of objects. The paper applies the OD on the k-dimensional dataset with k>=5. This approach uses the distance based outlier detection for multidimensional dataset. Clustering is process of a collection of data into groups with respect to a distance or similarity measure.

The objective of the clustering is to divide data into different groups by using their similarities. Data objects are added to the group with which its similarity is higher than the other groups. In data mining, clustering is used to discovery of the distribution of data and the detection of patterns. Here authors have proposed a new clustering algorithm called C2P. This approach exploits index structures along the processing of closest pair queries in spatial databases. It combines the advantages of the hierarchical agglomerative and graph-theoretic clustering algorithms.

II. LITERATURE SURVEY

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.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

In [3] Author proposes 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 proposes 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. OUTLIER DETECTION TECHNIQUES

The paper provides extension for large spatial databases and for outlier handling The outlier detection techniques operate in one of the three models are;

1) Supervised outlier Detection:

These techniques are trained in supervised mode and consider the availability of labeled instances for normal as well as outlier classes in a training dataset.

2) Semi-supervised outlier Detection:

This technique is trained in supervised mode and considers the availability of labeled instances for normal and do not require labels for the outlier class.

3) Unsupervised Outlier Detection:

These techniques operate in unsupervised mode do not require training data from any class. There are many more outlier detection techniques based on the nearest neighbor which considers that outlier object appears far from their nearest neighbor. Such methods base on a distance or similarity measure to search the neighbors with Euclidean distance. Numbers of neighbor-based OD methods include defining the outlier score of a point as the distance to its k^{th} nearest neighbor.

IV. REVERSE NEAREST NEIGHBOR

Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlines 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 reexamination for the outlier-detection task. In this light, it will revisit the ODIN method.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

V. HIGH DIMENSIONAL DATA

Despite the general impression that all points in a high-dimensional data set seem to become outliers, we show that unsupervised methods can detect outliers which are more pronounced in high dimensions, under the assumption that all (or most) data attributes are meaningful, i.e. not noisy.

HUBNESS

The phenomenon of Hubness was observed, which affects reverse nearest-neighbor counts, i.e. k-occurrences (the number of times point x appears among the k nearest neighbors of all other points in the Data Set).

VI. RELATION BETWEEN ANTIHUB AND OUTLIER

Based on the relation between ant hubs and outliers in high- and low-dimensional settings, in explore two ways of using k-occurrence information for expressing the outlines of points, starting with the method ODIN.

The proposed system the scope of our investigation is to examine: (1) point anomalies, i.e., individual points that can be considered as outliers without taking into account contextual or collective information, (2) unsupervised methods, and (3) methods that assign an "outlier score" to each point, producing as output a list of outliers ranked by their scores. The most widely applied methods within the described scope are approaches based on nearest neighbors, which assume that outliers appear far from their closest neighbors. Such methods rely on a distance or similarity measure to find the neighbors, with Euclidean distance being the most popular option. Variants of neighbor-based methods include defining the outlier score of a point as the distance to its kth nearest neighbor. The angle-based outlier detection (ABOD) technique detects outliers in high-dimensional data by considering the variances of a measure over angles between the difference vectors of data objects.

The discussion of problems relevant to unsupervised outlier-detection methods in high dimensional data by identifying seven issues in addition to distance concentration: noisy attributes definition of reference sets, bias (comparability) of scores, interpretation and contrast of scores, exponential search space, data-snooping bias, and Hubness. The reverse k-nearest neighbor count is defined to be the outlier score of a point in the proposed method ODIN, where a user-provided threshold parameter determines Whether a point is designated as an outlier or not. A method for detecting outliers based on reverse neighbors was briefly considered, judging that a point is an outlier if it has a zero k-occurrence count. The proposed method also does not explain the mechanism which creates points with low k-occurrences, and can be considered a special case of ODIN with the threshold set to 0. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data warrant their reexamination for the outlier-detection task.

A. PROBLEM IDENTIFICATION:

1) Cluster boundaries can be crossed, producing meaningless results of local outlier detection. How to determine optimal neighborhood size(s)?
2) Monotone function only one point should be analysis and detect.
3) Investigate secondary measures of distance/similarity, such as shared-neighbor distances.

B. PROBLEM ANALYSIS:

High values of k can be useful, but: Computational complexity is raised; approximate NN search/indexing methods do not work anymore. Is it possible to solve this for large k?

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

C. PROBLEM SOLUTION:

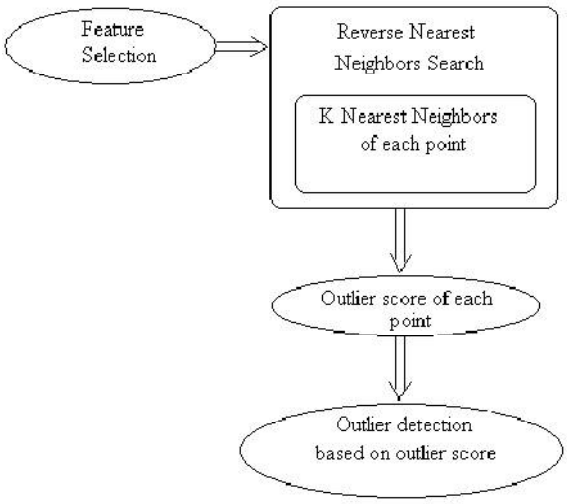
RNN SEARCH:

Reverse nearest neighbor (RNN) queries are a popular variant of RNN search. Capable of analysis two identical points at same time. Optimal region may contain an infinite number of points, how to represent and find such an optimal region become challenging in terms of running time. But using Max segment algorithm 100,000 times faster.

MAX SEGMENT ALGORITHM

The major reason why this algorithm is efficient is that we transform the optimal region search problem in a two-dimensional space to the optimal interval search problem in a one-dimensional space whose search space is significantly smaller than the search space in the two-dimensional space. After the transformation, it can use a plane sweep-like method to find the optimal interval efficiently. Finally, the optimal interval can be used to find the optimal region in the original two-dimensional space.

VII. ARCHITECTURE



The below figure shows that the overview of the proposed system. Pre-processing is done for eliminating missing values and making the data ready for the proposed work is done. Distance calculation is by default is a Euclidean Distance. This distance is for finding neighbor among the data points. The distance is an input for knn classification. The knn classification results in the list k-nearest neighbor for all points in the dataset. Among the list of k-nearest neighbor we find some points occur rarely and some points occur frequently known as antihub points and hub points respectively. Then finally outlier score is calculated and the outlier points are identified.

The outlier detection is done by the following methodologies for many real datasets.

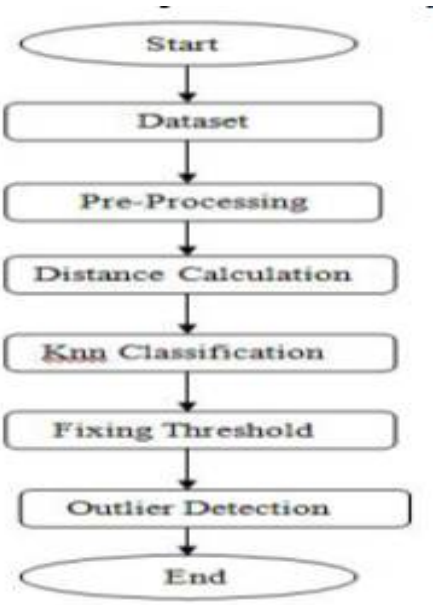
1. Euclidean distance
2. Calculating Nk(x)
3. Outlier score and performance analysis

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016



K-NN ALGORITHM

K-nearest neighbor algorithm is one of the classic data mining methods and its role is classification and similarity search. The k-nn returns the all objects that are match the query object in data set D. The below diagram represents the k-nn algorithm.

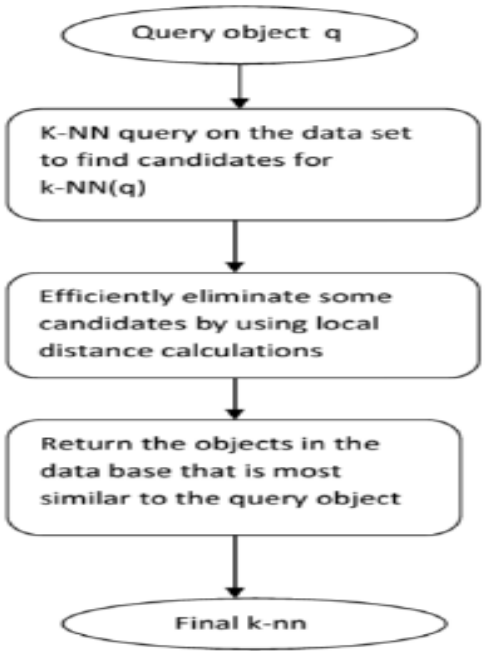


Fig: KNN algorithm

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

In the above diagram query object is given to the system ,the k-nn algorithm searches the all objects that are similar to the given query object and efficiently eliminate the some objects with the use of local distance calculations. The algorithm returns the objects the database that is most similar to the query object.

K-OCCURRENCES

K-occurrences identifying the number of occurrences in data set or database. It is a frequent item set mining, it identifying the frequent individual items In dataset and databases this are identified the frequent objects. The frequent objects are identified by using the appriori algorithm ,we have to calculate the confidence and support for the infrequent items in a dataset. We have to maintain the high confidence and minimum support.

1. Support: The X=>Y hold with support if the number of transactions or items in dataset D contains x+y. Support= occurrence/total support. Where occurrence is number of items found in frequent and total support is total number of transactions or related items.

2. Confidence: The X=Y holds with certainty (or) confidence if the number of transactions (or) items in dataset D that contains X likewise contain Y. Confidence=support(x+y)/support(x).

HUBNESS PHENOMENON

Nk(x), the number of k-occurrences of point x€ Rd, is the number of times x occurs among k nearest neighbors of all other points in a data set. In other words: Nk(x) is the reverse k-nearest neighbor count of x .Nk(x) is the in-degree of node x in the kNN digraph. Concentration of distance / similarity. High-dimensional data points approximately lie on a sphere centered at any fixed point.

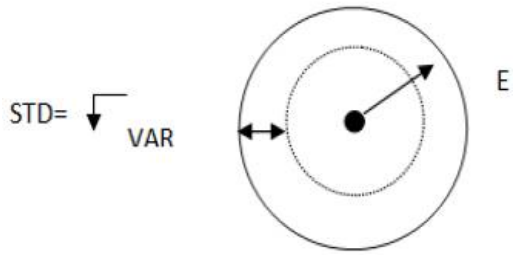


Fig: Hubness Phenomenon

Anti-Hubs in Outlier Detection:

In high dimensions, points with low Nk the anti-hubs can be considered distance based outliers. They are far away from other points in the data set / their cluster High dimensionality contributes to their existence. As dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods. The opposite can take place even when no true outliers exist, in the sense of originating from a different distribution this suggests that high dimensionality affects outlier scores and (anti-)Hubness in similar ways. Notable weakness of Antihub, discrimination of scores, contributed to by two factors: 1.Hubness 2.Discreteness of scores. So proposed method AntiHub2, which combines the Nk score of a point with Nk scores of its k nearest neighbors, in order to maximize discrimination. AntiHub2 improves discrimination of scores compared to the Antihub method.

1) Local outlier factor (LOF).
2) RNN Search.
3) Performance evaluation and result visualization.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

VIII. CONCLUSION

It provided a unifying view of the role of reverse nearest neighbor counts in unsupervised outlier detection. Effects of high dimensionality on unsupervised outlier-detection methods and Hubness .Extension of previous examinations of (anti)Hubness to large values of k the paper also explores the relationship between Hubness and data scarcity. It formulated the Antihub method, discussed its properties, and improved it in AntiHub2 by focusing on discrimination of scores Our main hope: clearing the picture of the interplay between types of outliers and properties of data, filling a gap in understanding which may have so far hindered the widespread use of reverse neighbor methods in unsupervised outlier detection.

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.