RNN (Reverse Nearest Neighbour) in Unproven Reserve Based Outlier Discovery

S. Priya
Assistant Professor
Department of Information Technology
Ganadipathy Tulsi's Jain Engineering College, Vellore, Tamilnadu, India

M. Srinivasan
Associated Professor
Department of Information Technology
Priyadarshini Engineering College, Vellore, Tamilnadu, India

Abstract

Outlier detection refers to task of identifying patterns. They don't conform establish regular behavior. Outlier detection in high-dimensional data presents various challenges resulting from the "curse of dimensionality". The current view is that distance concentration that is tendency of distances in high-dimensional data to become in discernible making distance-based methods label all points as almost equally good outliers. This paper provides evidence by demonstrating the distance based method can produce more contrasting outlier in high dimensional setting. The high dimensional can have a different impact, by reexamining the notion of reverse nearest neighbors. It is observed the distribution of point reverse count become skewed in high dimensional which resulting in the phenomenon known as Hubness. This provide insight into how some points (anti hubs) appear very infrequently ink-NN lists of other points, and explain the connection between anti hubs, outliers, and existing unsupervised outlier-detection methods. It crucial to understand increasing dimensionality so than have searching is different using maximum segment algorithm. Optimal interval search problem in a one dimensional space whose search space is significantly smaller than search space in two dimensional spaces.

Keywords- Outlier Detection, Reverse nearest Neighbours, High-Dimensional Data, Distance Concentration

I. INTRODUCTION

A. Objective:

This paper is to analysis high dimensional space in order to detect duplication data in unsupervised method. Further evidence reexamining method produced meaningful information. Easily searching and indexing neighbor size.

B. Scope of the Work:

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. Produce collective information of outlier score and get meaningful duplication data.

II. OVERVIEW OF OUTLIER DETECTION

Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice. The interest in outliers is strong since they may constitute critical and actionable information in various domains, such as intrusion and fraud detection, and medical diagnosis.

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. 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. This somewhat simplified view was recently challenged.

A. 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.

B. 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.

C. 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).

D. 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.

III. EXISTING METHOD

Nearest neighbor search and many other numerical data analysis tools most often rely on the use of the Euclidean distance. When data are high dimensional, however, the Euclidean distances seem to concentrate; all distances between pairs of data elements seem to be very similar. This paper justifies the use of alternative distances to fight concentration by showing that the concentration is indeed an intrinsic property of the distances and not an artifact from a finite sample.

In this paper presents the first algorithms for efficient RNN search in generic metric spaces. Our techniques require no detailed representations of objects, and can be applied as long as their mutual distances can be computed and the distance metric satisfies the triangle inequality.

A novel perspective on the problem of clustering high-dimensional data. Instead of attempt to avoid the curse of dimensionality by observing a lower dimensional feature subspace, we embrace dimensionality by taking advantage of inherently high-dimensional phenomenon. The tendency of high-dimensional data to contain points (hubs) that frequently occur in k nearest-neighbor lists of other points can be successfully exploited in clustering.

In this paper, we propose a novel approach named ABOD (Angle-Based Outlier Detection) and some variants assessing the variance in the angles between the difference vectors of a point to the other points. This way, the effects of the "curse of dimensionality" are alleviated compared to purely distance-based approaches. The idea of using the k nearest neighbors already resembles density based approaches that consider ratios between the local density around an object and the local density around its neighboring object.

High-dimensional data in Euclidean space pose special challenges to data mining algorithms. These challenges are often indiscriminately subsumed under the term 'curse of dimensionality', more concrete aspects being the so-called 'distance concentration effect', the presence of irrelevant attributes concealing relevant information, or simply efficiency issues. In about just the last few years, the task of unsupervised outlier detection has found new specialized solutions for tackling high-dimensional data in Euclidean space. These approaches fall under mainly two categories, namely considering or not considering subspaces (subsets of attributes) for the definition of outliers. The former are specifically addressing the presence of irrelevant attributes, the latter do consider the presence of irrelevant attributes implicitly at best but are more concerned with general issues of efficiency and effectiveness.

A. Drawbacks

1) Threshold value is used to differentiate outliers from normal object and lower outlines threshold value will result in high false negative rate for outlier detection.
2) Problem arises when data instance is located between two clusters, the inter distance between the object of k nearest neighborhood increases when the denominator value increases leads to high false positive rate.
3) Needs to improve to compute outlier detection speed.
4) Needs to improve the efficiency of density based outlier detection.

IV. PROPOSED METHOD

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?

C. Problem Solution:

1) Bi Chromatic RNN Search:

Bi chromatic reverse nearest neighbor (BRNN) 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.

V. ARCHITECTURE

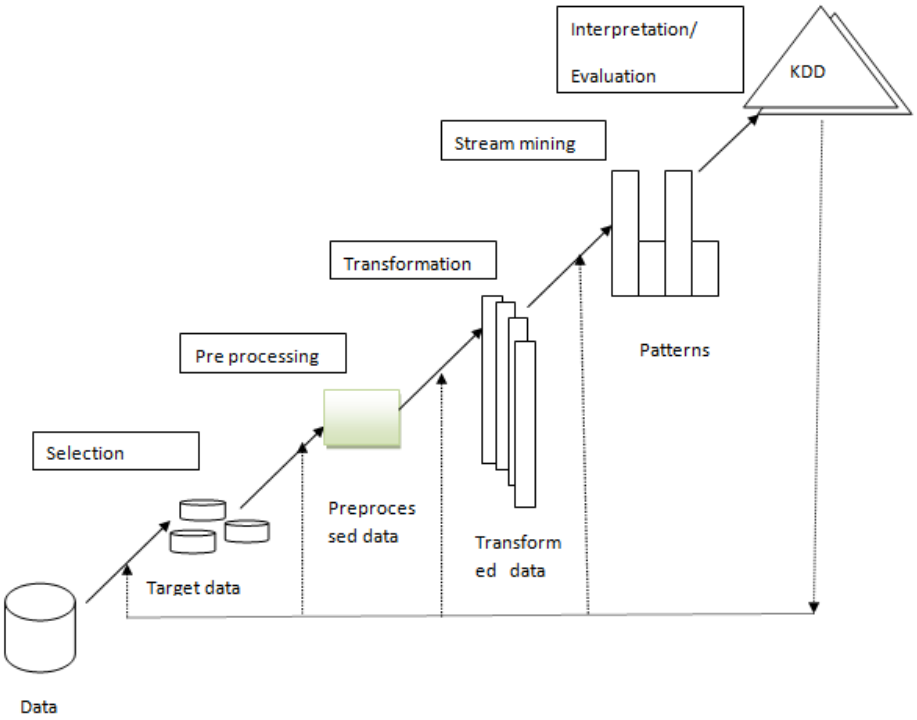

Fig. 1: Architecture

These are the modules for the efficient malicious node detection

1) Data preprocessing.
2) Data transformation.
3) Outlier detection in high dimensional.
- ABOD
- BRNN search

4) Anti-Hubs method.
- Relationship between Antihub and outlier
- Multimodality and neighborhood size
- Hubness phenomenon

5) Outlier detection methods based on Antihub
- Antihub
- Antihub2
- BRNN Search

6) Density based INFLO
7) Data post processing.
- Pattern evaluation
- Pattern selection
- Pattern interpretation

VI. DATA PREPROCESSING

A preprocessor is a program that processes its input data to produce output that is used as input to another program. The output is said to be a preprocessed form of the input data, which is often used by some subsequent programs like compilers.

A. Data Transformation:

Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. Data are transformed or consolidate into forms appropriate for mining by performing summary or aggregation operations, for instance.

B. Outlier Detection in High Dimensional:

In high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases. As dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods, assuming all dimensions carry useful information.

1) Angle-based outlier detection (ABOD)
2) BRNN Search

C. Influenced Outlines Measure (INFLO):

Based on a symmetric relationship that considers both neighbors and reverse neighbors of a point when estimating its density distribution. INFLO is essentially a density-based technique. It is used to design to work in settings of low to moderate dimensionality. The main focus of was on the efficiency of computing INFLO scores.

D. Angle-Based Outlier Detection (ABOD):

It detects outliers in high-dimensional data by considering the variances of a measure over angles between the difference vectors of data objects. ABOD uses the properties of the variances to actually take advantage of high dimensionality and appear to be less sensitive to the increasing dimensionality of a data set than classic distance-based methods

E. Bi Chromatic RNN Search:

Bi chromatic reverse nearest neighbor (BRNN) queries are a popular variant of RNN search. Given a point data set P, a site data set T, and a point q, the output of a BRNN query is {p€P dist(p, q) < NNdist(p, T), where NNdist(p, T) is the distance from p to its NN in T.

F. Anti-Hubs Method:

Antihub is a direct consequence of high dimensionality when neighborhood size k is small compared to the size of the data. Distance concentration refers to the tendency of distances in high-dimensional data to become almost indiscernible as dimensionality increases, and is usually expressed through a ratio of a notion of spread (e.g., standard deviation) and magnitude (e.g., the expected value) of the distribution of distances of all points in a data set to some reference point. If this ratio tends to 0 as dimensionality goes to infinity, it is said that distances concentrate.

1) The relation between Antihub and outliers.

2) Multimodality and neighborhood size
3) Hubness phenomenon NK

G. 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.

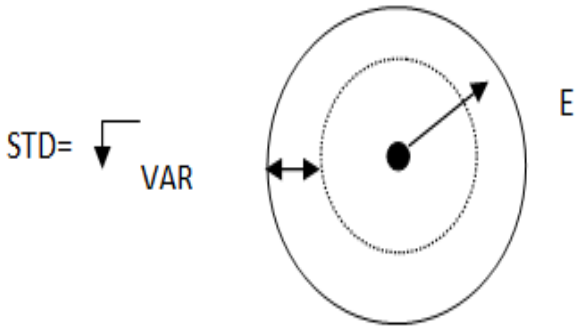

Fig. 2: Hubness Phenomenon

H. 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) Bi chromatic RNN Search.
3) Performance evaluation and result visualization.

I. Local Outlier Factor (LOF):

This is used to aggregate ratios of local reach ability distances.

J. Bi chromatic RNN Search:

Bi chromatic reverse nearest neighbor (BRNN) queries are a popular variant of RNN search. Given a point data set P, a site data set T, and a point q, the output of a BRNN query is {p€P dist(p, q) < NNdist(p, T), where NNdist(p, T) is the distance from p to its NN in T.

K. 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

VII. 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 sparsity. 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] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic, "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection", IEEE transactions on knowledge and data engineering, vol. 27, no. 5, may 2015.
[2] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7, pp. 873–886, Jul. 2007.
[3] Y. Tao, M. L. Yiu, and N. Mamoulis, "Reverse nearest neighbor search in metric spaces," IEEE Trans. Knowl. Data Eng., vol. 18, no. 9, pp. 1239–1252, Sep. 2006.
[4] N. Toma_sev, M. Radovanovi_c, D. Mladeni_c, and M. Ivanovi_c,"The role of hubness in clustering high-dimensional data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 739–751, Mar. 2014.
[5] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlierdetection in high-dimensional data," in Proc 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.
[6] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 363–387, 2012.