Unsupervised Distance-Based Outlier Detection in Reverse nearest Neighbours

Nageswara Rao A

Department of CSE, Eswar College of Engineering, Narasaraopet, JNTUK, AP, India

Abstract—Outlier detection in high-dimensional data presents various challenges resulting from the “curse of dimensionality.” A prevailing view is that distance concentration, i.e., the tendency of distances in high-dimensional data to become indiscernible, hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper, we provide evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, we show that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbours in the unsupervised outlier-detection context. Namely, it was recently observed that the distribution of points’ reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as hubness. We provide insight into how some points (anti hubs) appear very infrequently in k-NN lists of other points, and explain the connection between anti hubs, outliers, and existing unsupervised outlier-detection methods. By evaluating the classic k-NN method, the angle-based technique designed for high-dimensional data, the density-based local outlier factor and influenced outlierness methods, and anti hub-based methods on various synthetic and real-world data sets, we offer novel insight into the usefulness of reverse neighbor counts in unsupervised outlier detection.

I. INTRODUCTION

Outlier (anomaly) detection refers to the task of identifying patterns that do not conform to established regular behavior. 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., pairwise distances become indiscernible as dimensionality increases. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good outlier. This somewhat simplified view was recently challenged.

Our motivation is based on the following factors:

1) It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in the actual challenges posed by the “curse of dimensionality” differs from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space. We will present further evidence which challenges this view, motivating the (re)examination of methods.

2) Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness 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, we will revisit the ODIN method.

Our contributions can be summarized as follows:

1) We discuss the challenges that unsupervised outlier detection faces in high-dimensional space. 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. Our findings complement the observations from by demonstrating such behavior on data originating from a single distribution without outliers generated by a different mechanism. Also, we explain how high dimensionality causes such pronounced outlierness in comparison with low-dimensional settings.

2) Recently, 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). Hubness is manifested with the increase of the (intrinsic)
dimensionality of data, causing the distribution of k-occurrences to become skewed, also having increased variance. As a consequence, some points (hubs) very frequently become members of k-NN lists and, at the same time, some other points (antihubs) become infrequent neighbors. We examine the emergence of antihubs and the way it relates to outlierness of points, also considering low dimensional settings, extending our view to the full range of neighborhood sizes, and exploring the interaction of hubness and data sparsity.

3) Based on the relation between antihubs and outliers in high- and low-dimensional settings, in Section 5 we explore two ways of using k-occurrence information for expressing the outlierness of points, starting with the method ODIN proposed in. Our main goal is to provide insight into the behavior of k-occurrence counts in different realistic scenarios (high and low dimensionality, multimodality of data), that would assist researchers and practitioners in using reverse neighbor information in a less ad-hoc fashion.

4) We describe experiments with synthetic and real data sets, the results of which illustrate the impact of factors such as dimensionality, cluster density and anti-hubs on outlier detection, demonstrating the benefits of the methods, and the conditions in which the benefits are expected.

II. EXISTING SYSTEMS

A. Local outlier factor (LOF)

In LOF, compare the local density of a instances with the densities of its neighborhood instances and then assign anomaly score to given data instance. For any data instance to be normal not as an outlier, LOF score equal to ratio of average local density of k nearest neighbor of instance and local density of data instance itself. To find local density for data instance, find radius of small hyper sphere centered at the data instance. The local density for instances is computed by dividing volume of k, i.e k nearest neighbor and volume of hyper sphere. In this assign a degree to each object to being an outlier known as local outlier factor. Depends on the degree it determines how the object is isolated with respect to surrounding neighborhood. The instances lying in dense region are normal instances, if their local density is similar to their neighbors, the instances are outlier if there local density lower than its nearest neighbor. LOF is more reliable with top-n manner. Hence it is called as top-n LOF means instances with highest LOF values consider as outliers.

B. Local distance based outlier factor (LDOF)

Local distance based outlier factor Measure the objects outlierness in scattered datasets. In this uses the relative location of an object to its neighbors to determine the object deviation degree from its neighborhood instances. In this scattered neighborhood is considered. Higher deviation in degree data instance has, more likely data instance as an outlier. In this algorithm calculates the local distance based outlier factor for each object and then sort and ranks the n objects having highest LDOF value. The first n objects with highest LDOF values are consider as an outlier.

C. Influenced Outlierness (INFLO)

This algorithm considers the circumstances when outliers are in the location where neighborhood density distributions are significantly different, for example, in the case of objects close to a denser cluster from a sparse cluster, this may give wrong result. This algorithm considers the symmetric neighborhood relationship. In this considering influence space and when estimating its density distribution also considers both neighbors and reverse neighbors of an object. Assign each object in a database a influenced outlierness degree. The higher inflo means that the object is an outlier.

D. Disadvantages:

- Threshold value is used to differentiate outliers from normal object and lower outlierness threshold value will result in high false negative rate for outlier detection.
- 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.
- Needs to improve to compute outlier detection speed.
- Needs to improve the efficiency of density based outlier detection.

III. PROPOSED SYSTEM

It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in the actual challenges posed by the “curse of dimensionality” differs from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space. We will present further evidence which challenges this view, motivating the (re)examination of methods.

Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness 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, we will revisit the ODIN method.
A. Advantages of Proposed System

Demonstration of one plausible scenario where the methods based on antihubs are expected to perform well, which is in a setting involving clusters of different densities. For this reason, we use synthetic data in order to control data density and dimensionality.

B. System Architecture

![System Architecture Diagram]

1) Data collection and data pre-processing: In data collection the initial input data for this system will be collected from standard dataset portal i.e. UCI data set repository. As proposed in system, the standard dataset will be used for this system includes Cover type, IPS datasets. Collected datasets may be available in their original, uncompressed form therefore; it is required to preprocess such data before forwarding for future steps. To preprocess large dataset contents, techniques available is data mining such as data integration, data transformation, data cleaning, etc. will be used and cleaned, required data will be generated.

2) Data partitioning: In this module, as stated earlier in system execution plan, the preprocessed data is divided into number of clients from central supervisor node i.e. server as per the data request made by desired number of clients. This partitioned data will be then processed by individual clients to identify outliers based on applied algorithm strategy.

3) Outlier detection: The technique proposed for identifying outliers will be applied initially at distributed clients and their results of detected outliers would be integrated on server machine at final stage computation of outliers. To do this, the outlier detection strategies proposed are KNN Algorithm with ABOD and INFLO Method.

The proposed algorithm considers the k-occurrences defined as dataset with finite set of n points and for a given point x in a dataset, denote the number of k-occurrences based on given similarity or distance measure as Nk(x), that the number of times x occurs among all other points in k nearest neighbor and points those frequently occurred as a hubs and points those occur infrequently as an antihub. Uses reverse nearest neighbors for instance, finding the instances to which query object is nearest. In this first read the each attribute in high dimensional dataset, then using angle based outlier detection technique compute the distance for every attribute using dataset Set distance and compare with distance from each instance and assign the outlier score. Based on that outlier score using reverse nearest neighbor determine that particular instance is an outlier or not.

4) 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 prop.

IV. 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


