International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Maximizing Biochromatic Reverse Nearest Neighbors In Unsupervised Outlier Detection.

R .Iswariya [1]mtech(it).
Mr.D.Duraikumar[2] (HOD) Associate professor
Prof S.Priya[3] Assistant professor
Ganadipathy Tulsi's Jain Engineering College Vellore

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 max segment algorithm. Optimal interval search problem in a one dimensional space whose search space is significantly smaller than search space in two dimensional space. Maximizing Biochromatic reverse search. In transform the optimal region search problem in a global data object to optimal interval search problem in a local data object search space efficient to global data object using maxsegment algorithm. Further evidence reexamining method produced meaningful information.

***Index Terms*— Outlier detection, reverse nearest neighbors, high dimensional, maxsegment,Biochromatic.**

INTRODUCTION

Outlier detection is to analysis high dimensional space in order to detect duplication data in unsupervised method. 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. 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 a representative alabels 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

Reverse Nearest Neighbor

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, it will revisit the ODIN method.

High Dimensional Data

Revisit the commonly accepted view that in high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases. In this view was challenged by showing that the exact opposite may take place: 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.

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

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. K-occurrences (the number of times point x appears among the k nearest neighbors of all other points in the Data Set).

Relation between Antihub and Outlier

Their relation to outliers detected by unsupervised methods in the context of varying neighborhood size k.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 outlierness of points, starting with the method ODIN. Two ways of using k-occurrence information for expressing the outlierness of points, starting with the method ODIN proposed. 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).

Existing system.

[1] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," Nearest neighbor searc h 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. Furthermore, an estimation of the concentration as a function of the exponent of the distance and of the distribution of the data is given.

[2] Y. Tao, M. L. Yiu, and N. Mamoulis, "Reverse nearest neighbor search in metric spaces," In this project 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.

[3] N. Toma_sev, M. Radovanovi_c, D. Mladeni_c, and M. Ivanovi_c,"The role of hubness in clustering high-dimensional dataA 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

[4] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," In high-dimensional data, these approaches are bound to deteriorate due to the notorious "curse of dimensionality". 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.

[5] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data,". 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.

Drawbacks:

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.

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.

PROPOSED SYSTEM

The proposed system the scope of our investigation is to examine: (1) Anomalies, i.e. considered as outliers taking into account contextual or collective information, (2) unsupervised methods, and 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.Bichromatic reverse nearest neighbor (BRNN) queries are a popular variant of RNN search. Max segment algorithm: The major reason why this algorithm is efficient is that transform the optimal region search problem in global to the optimal interval search problem in a local space whose search space is significantly smaller than the search space in the global. After the transformation, it can use a plane sweep-like method to find the optimal interval efficiently searching.

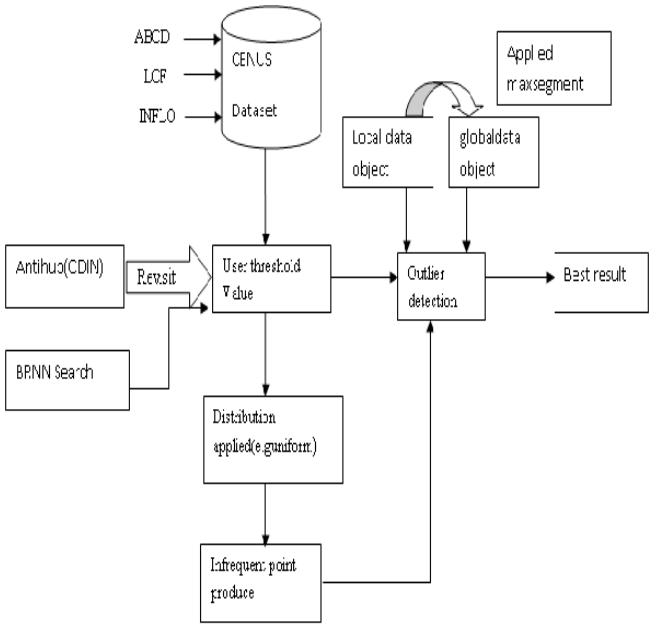


Fig:1. Antihub calculation process

Antihubs

It provided revisit the commonly accepted view that in high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases. Let applied setting numbers of drawn points, and distance measure.The demonstrated behavior is actually an inherent consequence of increasing dimensionality of data, with the tendency of the detected prominent outliers to come from the set of antihubs points that appear in Very few, if any, nearest neighbor lists of other points in the data.In examine fastABOD method variant assign outlier score is not expected anomaly detection.

Antihub algorithm:

Algorithm 1 AntiHub$_{dist}(D,k)$ (based on ODIN)

Input:

- Distance measure *dist*
- Ordered data set $D = (x_1, x_2, \ldots, x_n)$, where $x_i \in \mathbb{R}^d$, for $i \in \{1,2,\ldots,n\}$
- No. of neighbors $k \in \{1,2,\ldots\}$

Output:

- Vector $s = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n$, where s_i is the outlier score of x_i, for $i \in \{1,2,\ldots,n\}$

Temporary variables:

- $t \in \mathbb{R}$

Steps:

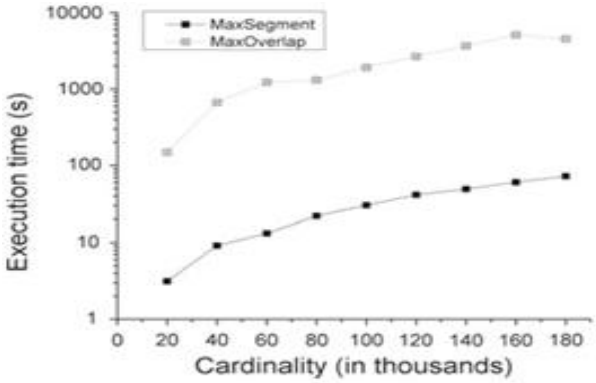
1) For each $i \in (1,2,\ldots,n)$
2) $t := N_k(x_i)$ computed w.r.t. *dist* and data set $D \backslash x_i$
3) $s_i := f(t)$, where $f : \mathbb{R} \to \mathbb{R}$ is a monotone function

MaxSegment algorithm

```
Algorithm MaxSegment algorithm
1: // Phase 1
2: for each user point o ? O do
3: search the nearest neighbor of o in P, says p
4: construct an DS c, centered at o with radius |p,α|
5: end for
6: // Phase 2
7: choose the DS c with the largest w(c)
8: initialize Max Inf ? w(c) and MaxS ? {c}
9: for i = 1 to |O| do
10: // Step 1
11: find all s intersected with NLC ci and store them into list L
12: // Step 2
13: for each NLC cj ? L do
14: generate intersection arc e = (q1, q2) where q1 and q2 are the
intersection points between the boundaries of ci and cj, and assign both
q1.DS and q2.DS with cj
15: if q1 > q2 then
16: generate two sub-intersection arcs e1 = (q1, 360?) and e2 = (0?,q2)
17: end if
18: store intersection points of the generated intersection arcs into Q
19: end for
20: // Step 3
21: sort the intersection points in Q according to their angle values
22: initialize Inf ? w(ci), S ? {ci}
23: // Step 4
24: for each intersection point t ? Q do
```

RESULT:

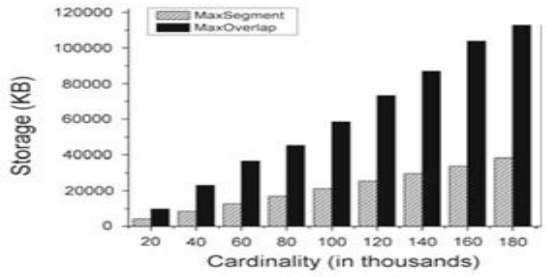
The goal of our experimental evaluation is to examine the effectiveness of outlier mining methods described in the previous section. Its second objective is to examine the behavior of the methods with respect to the k parameter. The main overall aim of this section is to further support the observations that reverse-neighbor relations can be effectively applied to outlier detection in both high- and low dimensional settings.



(a) Execution time

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.



(b) Storage

Extension of previous examinations of (anti)hubness to large values of *k.* It formulated the AntiHub method, discussed its properties, and improved it in AntiHub² by focusing on discrimination of scores. The major reason why this algorithm is efficient is that we transform the optimal region search problem in a local data space to the optimal interval search problem in a global data space whose search space is significantly smaller than the search space in the space

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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

[5] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist. Anal. DataMining,2012