Outlier Detection using Semi Supervised Data with Reverse Nearest Neighbors

A Somakala
M.Tech Student, Department of CSE
JNTUA College of Engineering, Ananthapuramu
Andhra Pradesh, India

P Dileep Kumar Reddy
Lecturer, Department of CSE
JNTUA College of Engineering, Ananthapuramu
Andhra Pradesh, India

***Abstract*—Outlier Detection is very prominent in determining characters and features of a given data and detecting it high dimensional has its own challenges like Curse of Dimensionality. It was a common notion that distance based methods was not ideal in detecting outliers, in high dimensional data. Recently Reverse Nearest Neighbors counts, where, to how many points, a point is a neighbor is calculated, gave a new direction to detect outliers in high dimension giving rise to concept called hubness.In any given dataset few points has more information than other points, by virtue these points occur less frequently than other points. They can be determined with the help of hubness phenomenon. These data points will be the point of interest in this paper. Supervised learning is to classify the data by assigning labels to input data. By learning from the past occurrencesand assigning label to the outcome, we will determine the semi supervised outliers with reverse neighbors, there by overcoming the hubness aware approaches for classified machine learning in high dimensional data.**

Keywords—Reverse nearest neighbors, Hubness, Hubs, Anti-hubs, High dimensional data, Outliers

I. INTRODUCTION

Detection of outliers in a dataset is to identify the data points that differ from the overall data points. Ex: Determining star football players from a group of football players. Accuracy of detecting these outliers has a great importance in various fields such as medical, forensics, fraud detection, hacking etc., depending upon the existence of labels the detection of outliers is classified as unsupervised, semi-supervised and supervised.Semi-supervised, supervisedand unsupervised are generally most fundamental machine learning tasks.

Data management and learning from the data sets is becoming prominent in day to day life, with the total amount of data around us is increasing on daily basis. The hugeexplosion of very large databases in today's companies and scientific institutions operates on high dimensional data, hence machine learning for the concerned in that field. High dimensional [1][3] data in turn faces challenges such as—the curse of dimensionality. One of the challenges in high dimensional data is (1) distance concentration, where machine learning techniques are directly affected. This is a concept where in high-dimensional data distances between all pairs of points to become almost equal making every point an outline.The other challenge of curse of dimensionality is (2) hubness [12][1]inherits nearest neighbor methods. Reverse nearest neighbor algorithm is used to generate hubs. By exploiting this phenomenon we can learn from past occurrences from there we can determine outliers with semi-supervised learning.

Reverse nearest neighbor algorithm [1] is where we calculate, let x be a point k be the number of occurrences then Nk is the count of how many times x is among k nearest neighbors for every point in a given dataset. Nk is the neighborhood score.For all the points in the dataset the data points with highest Nk count belong to a hub. In any given dataset few points has more information than other points, by virtue these points occur less frequently than other points. These data points will be the point of interest in this paper.

II. RELATED WORK:

Miloˇ sRadovanovi´c et al [1] in their work they detected outliers for unsupervised data with reverse nearest neighbors using ODIN method. They have proposed a unifying view of the role of reverse nearest neighbor counts in unsupervised outlier detection of how unsupervised outlier detection methods are affected with the higher dimensionality. These parameters are extended for large values of k. Relationship between hubness and sparsity are explored. Mainly they cleared about how properties of data and type of outliers are interpreted. This helped in increase in reach of reverse nearest neighbors. In this paper we are improvising by doing classification with semi supervised data.

Classification of a high dimensional datais challenging, be it unsupervised, semi or supervised learning. A lot of and researches are ongoing to get best results from these machine learning techniques. Hubness is a recent development. Emergence of hubs and their influence in data retrieval and the skewness of the data and its various aspects in high dimensional data aresummarized byMiloˇ sRadovanovi´c et al.[7]Determining of neighbors to a data point can be done with k-nearest neighbors (kNN). [2] With this Nk can be calculated. This can be calculated with help of Euclidean distance. Depending upon density local distance-based outlier factor (LDOF) [13] and local outlier probabilities (LoOP) [8][9] methods were used to detect outliers. Angle-based outlier detection (ABOD) can be used in high dimensional data to detect outliers methods were used [10].

Hubness aware classification for getting information in supervised or semi supervised learninghubness fuzzy k-NN (h-FNN) [5] algorithms were used. This approach can be used to learn data from past occurrences.

III. PROPOSED WORK:

Outlier detection by giving labels to the input data using reverse nearest neighbor algorithm i.e., with semi supervised data learning has three major aspects (1) Point anomalies- without collective and contextual information every point in the dataset can be considered as an outlier (2) To assign count of number of neighbors using reverse nearest neighbors algorithm and (3) Hubness aware Semi supervised methods to train the data and to assign labels to input data. With all these methods implemented we will determine outliers. By giving labels to the input data more meaning to outliers can be obtained.

A. OUTLIERS IN HIGH DIMENSIONAL DATA

When data in high dimensional set is taken into account Curse of Dimensionality comes into picture. Hubness is the event that is observed in high dimensional data. Distance concentration is another one, where even when an outlier is not expected an outlier will be shown.

B. HUBS AND ANTIHUBS

Hubs and antihubs are newly developed concepts to overcome the challenges faced when classifying high dimensional data. Before getting into hubs it is better to know about count of number of neighbors also can said as neighborhood score Nk. Neighborhood score Nk can be defined as the number of times a point is a neighbor, to k nearest neighbors. Assuming every point is a neighbor to itself Nk is greater than or equal to one. By considering this undefined values can be removed.

In the dataset to any point if Nk is greater than a threshold value then, those points belong to hub. All other points are antihubs. Here we can say all the data points of a hub have similar features, whereas antihubs have different features than main theme of whole dataset. By studying the features of hubs and antihubs we can see that few data points have more information than others, there by being less frequently occurred events, which are antihubs.

C. INFORMATIVENESS AND OUTLIERS

In semi supervised classification learning from past occurrences [11]is needed. With the help of this informativenessa new data point, let it be ‘x’, is assigned to a hub or classified as an outline. For a given Nk

Let $P(x) = Nk/n$

Where n is number of data points in a data set.Then,informativeness

$I(x) = \log(1/P(x))$

Relative and Absolute informativeness can be used to calculate outliers in local and global context. Let Relative Informativenessbea(x) and absolute informativeness be b(x) where,

$a(x) = [I(x)-\min I(x)]/[\log(n) – \min I(x)]$

$b(x) = I(x)/\log(n)$

Finally probabilities are calculated from the information gathered to assign a label to a new data point (y) is,

$P(y = c |x)=Nk, c(x)/Nk(x)$

Depending upon the probabilities [6] a data point will be assigned to a class c.

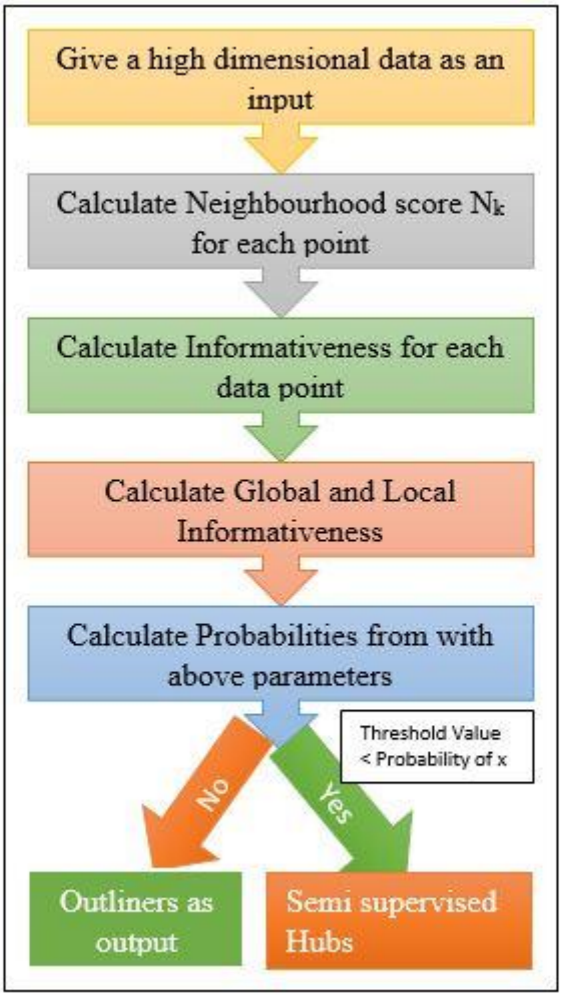

Give a high dimensional data as an input
Calculate Neighbourhood score N_k for each point
Calculate Informativeness for each data point
Calculate Global and Local Informativeness
Calculate Probabilities from with above parameters
Threshold Value < Probability of x
No
Yes
Outliners as output
Semi supervised Hubs

Figure1:Flow chart of the proposed work

IV. ALGORITHM FOR SEMI SUPERVISED DATA

In the process of determining outliers for semisupervised data we first do the following. (1) Get the count of neighbors of point i.e., neighborhood score. (2) With this score we determine to which class a point belong and assign its label depending upon the probabilities that are calculated.

A. GET NEIGHBORHOOD SCORE

k - Reverse nearest neighbors is used in getting the neighborhood score of a point (Nk)

Algorithm Part I:

For any Dataset D and k:

Input:

Let Distance be d

Data set D = (x1, x2… xn), where xi is a data point*i* belongs to {1,2,3..n}

No. of neighbors k belongs to {1,2,3….}

Output:

Vector NS = (ns1,ns2,…nsn) where ns is the neighborhood score for each data point.

Steps:

```
For everyi belongs to {1,2,3…n}
Nk(xi) = 0
for all distances < d
            Nk(xi)+ = Nk(xi)
      end for
end for
```

With above algorithm we calculate the neighborhood scores of any data point in the data set.

B. ASSIGN LABELS TO INPUT DATA

When assigning labels to input data, training data should be given as an input before classification. Hence to get training data, divide whole data set points into 10 parts and take one of the ten parts as a training data. Let that be T.

Algorithm Part II:

For any Data set D, k and NS

Input:

Ordered data set D = (x1, x2… xn), where xi is a data point i belongs to {1,2,3..n}

No. of neighbors k belongs to {1,2,3….}

Neighborhood scores NS = (ns1,ns2,…nsn) where ns is the neighborhood score for each data point.

Output:

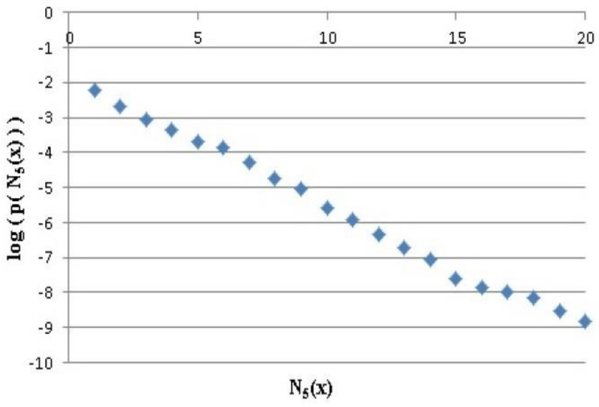
Labeled dataas an output for every data point.

Steps:

```
Calculate I(xi)for each I and NS
Calculate a(xi) and b(xi)
For all i
Calculate Pk(y = c |x)
      End for
```

With this algorithm labels are assigned to input data.

V. EXPERIMENTAL RESULTS:



The outliers are detected using reverse nearest algorithm with semi supervised data. These outliers will be more accurate when compared to using unsupervised data. Hubs and antihubs are created for the data sets. After calculating neighborhood score, we calculate the informativeness of any point. From there relative and absolute informativeness of a data point is calculated. With the help of relative and absolute informativeness for a data point probabilities of belong to a class are calculated.

Creation of hubs and antihubs are done in semi supervised classification. But outliers will not have any labels.

When plotted a graph across no. of neighbors and the informativeness of those points above graph is obtained.

VI. CONCLUSION AND FUTURE WORK:

Semi supervised data outliers detection is done with the help of k reverse nearest neighbor (KRNN). The concept of hubs is used to get neighborhood score. By exploiting the features of hubs and antihubs we are learning from past occurrences to get outliers. By using semi supervised learning we have increased accuracy of determining outliers than unsupervised learning. Relation between number of neighbors and the information in a data point is explored.

This can be further extended by examining few methods for supervised data. And best distance or similarity methods such as shared neighbor distances can be studied to get outlines. In aspect speed compared to accuracy can also be further exploited.

REFERENCES

[1] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovi"Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection" in IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 5, May 2015

[2] S. Ramaswamy et al. "Efficient algorithms for mining outliers from large data sets."SIGMOD Record, 29(2):427–438, 2000.

[3] M. Radovanović et al. Hubs in space: "Popular nearest neighbors in high-dimensional data". Journal of Machine Learning Research,11:2487–2531, 2010.

[4] N. Tomaˇsev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "Hubness-based fuzzymeasures for high dimensional k-nearest neighbor classification," in Machine Learningand Data Mining in Pattern Recognition, MLDM conference, 2011.

[5] "A probabilistic approach to nearest neighbor classification: Naive hubnessbayesian k-nearest neighbor," in Proceeding of the CIKM conference, 2011.

[6] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, "Nearest neighbors in highdimensionaldata: The emergence and influence of hubs," in Proc. 26th Int. Conf.on Machine Learning (ICML), 2009, pp. 865–872.

[7] C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distancemetrics in high dimensional spaces," in Proc. 8th Int. Conf. on Database Theory(ICDT), 2001, pp. 420–434.

[8] H.-P. Kriegel, P. Kr€oger, E. Schubert, and A. Zimek, "LoOP: Localoutlier probabilities," in Proc 18th ACM Conf. Inform. Knowl. Manage.,2009, pp. 1649–1652.

[9] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlierdetection in high-dimensional data," in Proc 14th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.

[10] N. Tomasev and D. Mladenic, "Nearest neighbor voting in highdimensional data: Learning from past occurrences," Comput. Sci. Inform.Syst., vol. 9, no. 2, pp. 691–712, 2012.

[11] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic,"The role of hubness in clustering high-dimensional data," IEEETrans. Knowl. Data Eng., vol. 26, no. 3, pp. 739–751, Mar. 2014.

[12] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlierdetection approach for scattered real-world data," in Proc 13thPacific-Asia Conf. Knowl. Discovery Data Mining, 2009, pp. 813–822.