An Enhanced Approach in Data Mining for Outlier Detection Based on Local Outlier Factor (LOF)

Dr. Deepak Dembla[1], Aditya Dixit [2], Sanjay Tiwari [3],
[1]Professor, Department of Computer Science & Engineering AIET,Jaipur, INDIA
[2] M. Tech. Research Scholar, Department of Computer Science & Engineering, AIET, Jaipur, INDIA
[3] Assoc. Prof , Department of Information technology ,AIET ,Jaipur, INDIA
deepak_dembla@yahoo.com[1] matforct@gmail.com[2], sanjaytiwari@aryaaiet.ac.in[3],

Abstract- The detection of outliers has recently become a very significant problem criterion in many commercial applications. This problem is further difficult by the fact that in many cases, the difference between outliers and normal data is very much unclear. In this paper, a modified approach for LOF (Local Outlier Factor) algorithm, appropriate for detecting outliers, is proposed. The proposed modified LOF algorithm provides equivalent detection performance as the basic LOF algorithm, while requiring approximately same or less computational time complexity. The paper provides Practical evidence that insertion of one factor z for the calculation of k-distance in the LOF will leads to an improvised algorithm in terms of no of outliers. Our experiments performed on real life data sets have demonstrated that the proposed LOF$_{ENH}$ algorithm is computationally same efficient as the original LOF, while at the same time is 64.25 percent more successful in detecting outliers on average as compared to LOF. LOF$_{ENH}$ can be very helpful in Network intrusion detection. LOF$_{ENH}$ algorithm is based on the z-distance and the optimum value of z for different types of problems will be a scope for future work.

1. INTRODUCTION

Outliers are also referred to as abnormalities, deviants, or anomalies in the data mining and statistics literature. In many applications, the data is found by one or more data generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights. In most of applications, the data has a “normal” model, and anomalies are recognized as deviations from this normal model. In many applications such as Credit Card Fraud, Interesting Sensor Events, Medical Diagnosis, Law Enforcement, Earth Science, the outliers can only be discovered as a sequence of multiple data points, rather than as an individual data point.

Hawkins defines the outlier as “an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism” [1]. From this definition, we can say that Outliers are deviant cases. Many good techniques on outlier mining exist. LOF (Local outlier Factor) formulation for outlier detection is considered to be quite efficient [2][3]. In this paper we are providing an improvement on LOF. This paper is organized into six sections. Section 2 is related with LOF. In section 4 we provide the definitions of enhanced versions of LOF. In section 5 evaluation results are given. Section 6 concludes the paper.

2. RELATED WORK (LOCAL OUTLIER FACTOR)

The concept of Local Outlier factor (LOF)[4] was first introduced by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander. Perhaps LOF introduced, first time the measure that quantifies how much an object is outlying with respect to other objects in a given database. Necessary definitions to explain LOF concept are given below. For further details [4] can be referred. In the following definition k is a user supplied natural number and is also referred known as minimum points MinPts. MinPts and k can be used interchangeably.

Definition 1: (k-distance of an object o denoted as k-Dist (o)):- it is the distance between object o and its k nearest neighbor.

Definition 2: (k-distance neighborhood of an object o denoted as N $_{k\text{-Dist(o)}}$ (o)):- Given the k-Dist(o), the N $_{k\text{-Dist(o)}}$ (o) is the set of number of objects having a same or less than the k-distance from o.

Definition 3: (reachability distance of o with respect to object p denoted by reach-Dist$_k$(o, p)):- it is the maximum distance out of k-Dist of an object p and real distance between o and p.

LOF is based on the concept of MinPts and reachability distance referenced from DBSCAN [5] and OPTICS[6]. In LOF formal definition MinPts is an indication of mass and the values reach - dist $_{MinPts}$ (o, p) for p □ N $_{MinPts}$ (o) , is an indication of volume to calculate the density in the neighborhood of an object o (lrd, as given in the next definition, is an indication of density).

Definition 4: (local reachability density of an object o):-It is denoted as lrd $_{Minpts}$ (o)):- The local reachability density of an object o can be calculated dividing one by the average reachability distance based on the MinPts-distance neighborhood of o[7][8][9].

Definition 5: (local outlier factor of an object o):- The LOF of o is defined as

$$LOF_{Minpts}(o) = \frac{\sum_{p \in N_{MinPts}(p)} \frac{lrd_{Minpts}(p)}{lrd_{Minpts}(o)}}{|N_{MinPts}(o)|}$$

3. PROBLEM STATEMENT

It is clear from the above discussion of LOF, that local reachability density (lrd) is an indication of density of the region surrounding the object point. But MinPts-dist also possesses the same indication for density. Small MinPts-dist maps to dense region i.e. high density while large MinPts-dist maps to sparse region (low density).The main objective of is to enhance the LOF so that it can detect a large number of correct outlier.

4. IMPROVEMENT ON LOF

LOF is a density based algorithm to identify outliers. The lrd is used as an indication of density which is based on MinPts-distance neighborhood and reachability distance. While reachability distance and MinPts-distance neighborhood are indication of volume and mass respectively [1]. These two parameters are dependent on MinPts-distance or k-distance if we consider k as MinPts. k-distance is defined in definition-1 as the distance between object o and its kth nearest neighbor. In our improvement we include a factor z to identify distance of an object o, (given k) denoted as z-distance in place of k-distance, as a distance from o to its k nearest objects with addition of z. Subsequent definitions including z-distance are as follows.

Definition 6: (z-distance of an object o): - For any positive integer k, the z-distance of an object o, denoted as z-distance (o), is defined as :

z-distance k (o) = $z+\sum_{p \in KNN(o)} d(o,p)$

We will use notation z-distance (o) in place of z-distance k(o). Here k is a natural number and z is some user supplied value to enhance the result.

Definition 7: (z-distance neighborhood of an object o):-For the given z-distance of o, the z-distance neighborhood of o contains every object whose distance from o is same or smaller than the z-distance, i.e.

$N_{z-distance(o)}$ (o) = {q □ D \{p} | d(o,q) ≤ z – distance(o) }

Here each object q is the z-nearest neighbors of o. When there is no confusion, we will use Nz(o) as a shorthand of $N_{z-distance(o)}$ (o)

Definition 8: (reachability distance of an object o with respect to object p):- The reachability distance of object o with respect to object p is defined as

reach - dist z (o) = max { z - distance(p), d(o, p)}

Definition 9: (local reachability density of an object o):-The local reachability density (lrd) of o is defined as

$$lrd\ z(o) = 1 \Bigg/ \frac{\sum_{p \in N_{z-distance(o)}(o)} reach - dist\ z\ (o, p)}{|N_{z-distance(o)}(o)|}$$

Definition 10: (local outlier factor of an object o):- The local outlier factor of an object o is defined as

$$LOF_{ENH}\ k\ (o) = \frac{\sum_{p \in N_{Z-distance(o)}(o)} \frac{lrd_{z}(p)}{lrd_{Z}(o)}}{|N_{z-distance(o)}(o)|}$$

Complexity analysis of LOF is well discussed in [9]. LOF_{ENH} is using the same methodology in terms of steps or passes. Therefore time complexity of LOF_{ENH} is same as LOF.

It is clear from formula that if lrd of object o is low and that of its neighborhood is high than $LOF_{ENH}(o)$ will be higher indicating it's higher outlier-ness.

5. RESULTS & ANALYSIS

In many of the experiments, synthetic dataset has been used to verify verifying performance of outlier mining algorithms. We evaluated our algorithm on real database downloaded from UCI Repository. Names are KDD CUP 99 Network Connections Data Set [10].

KDD CUP 99 Network Connections Data Set This data set was used in The Third International Knowledge Discovery and Data Mining Tools Competition. This competition was organized in collaboration with KDD-99. In competition the task was assigned to build a network intrusion detector. This detector is a predictive model that can distinguish between “bad" and good connections. Bad connection is called intrusions (or attacks), and “good" connections are normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. Attacks (or intrusions) can be divided into four main categories:

DOS: Denial-Of-Service, e.g. syn flood.
R2L: unauthorized access from a remote machine, e.g. guessing password.
U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow" attacks.
PROBING: Surveillance and and other probing, e.g.port scanning.
Details of database can be found in [10]. Number of Instances in our experiments are 1000 out of which there are 43 attack Instances.
We executed above algorithm for the above datasets and compared the results. Summary of the results are shown in Table 3. In the experiment we considered k = 10 and MinPts = 10 and z =1.

LOF

INPUT		OUTPUT	
MINPTS	**LOF Upper Bound**	**No Of Outlier Reported**	**No Of Correct Detections**
10	4.70848	10	6
10	3.66733	20	9
10	2.98519	30	13
10	2.78406	40	16
10	2.49591	50	18
10	2.21843	60	20
10	2.0761	70	22
10	1.97572	80	22
10	1.85329	90	25
10	1.75023	100	25

Table 1: Results of running LOF algorithm on KDD CUP-99 when z=1

LOF$_{ENH}$

INPUT		OUTPUT	
MINPTS	**LOF Upper Bound**	**No Of Outlier Reported**	**No Of Correct Detections**
10	1.8669826	10	10
10	1.6482171	20	18
10	1.4980196	30	25
10	1.3695992	40	29
10	1.2907759	50	31
10	1.2203705	60	32
10	1.1911992	70	34
10	1.1102141	80	36
10	1.0871897	90	37
10	1.0709447	100	37

Table 2: Results of running LOF$_{ENH}$ algorithm on KDD CUP-99 when z =1

In the above experiment we considered k = 10 and MinPts = 10 and z =1. Comparison of result is shown in Table 4.3. Basis of comparison between LOF and LOF$_{ENH}$ is performance. Hear performance means no. of correct detection of outliers.

No Of Outlier Reported	No. of correct detection	
	LOF	LOF$_{ENH}$
10	6	10
20	9	18
30	13	25
40	16	29
50	18	31
60	20	32
70	22	34
80	22	36
90	25	37
100	25	37

Table 3: Comparison Results with k = 10 and MinPts = 10 and z = 1 for KDD CUP 99 Data Set for LOF and LOF$_{ENH}$

In Table 3, total number of outliers detected (including correct and incorrect) and actual detection made by LOF and LOF$_{ENH}$ is shown. Above table makes it clear that algorithm LOF$_{ENH}$ is performing well w.r.t. LOF. In figure.1, Graph for the above results is shown With the result of above experiment shows that LOF$_{ENH}$ is performing consistently well.

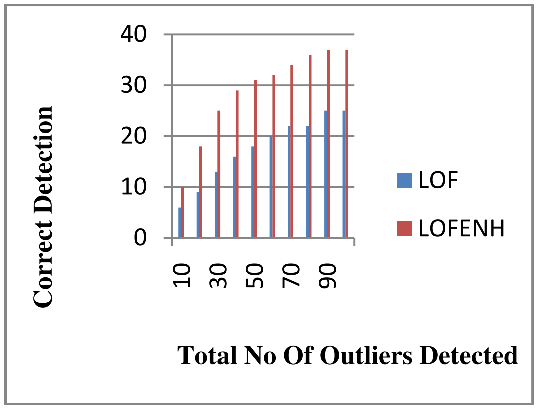

Figure 1: Actual Outliers detected by LOF$_{ENH}$ and LOF

6. CONCLUSIONS AND FUTURE SCOPE

In this research, we proposed the modification in the algorithm of LOF. This modification is quite useful. By replacing k-distance with z-distance. We noticed that significant improvement occurred in the performance of LOF. However no change in time complexity was found. The biggest issue in outlier detection is to detect the maximum number of correct outliers in a given dataset. Sometimes the outliers detected by an algorithm are not correct or we can say that it detect fake or wrong outliers that's why there is always a need of developing an algorithm which can detect maximum number of outliers correctly. Also the algorithm should be efficient both in terms of time and space complexity. This algorithm is based on the factor z and the optimum value of z for different types of problems will be a scope for future work.

8. REFERENCES

[1] D.Hawkins, “Identification of Outliers”, Chapman and Hall, London, 1980.

[2] J. Han and M. Kamber, “Data Mining, Concepts and Techniques”,Morgan Kaufmann, San Francisco, 2001.

[3] Aggarwal, C. C., Yu, S. P., “An effective and efficient algorithm for high-dimensional outlier detection”, The VLDB Journal, vol. 14, pp.211-221, 2005.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander,“LOF:Identifying density-based local outliers”, Proc. ACM SIGMOD International Conference on Management of Data, pp. 93–104, 2000.

[5] M. Ankerst,M. M. Breunig, H.-P.Kriegel, and J. Sander, “OPTICS:Ordering Points To Identify the Clustering Structure”, Proc. ACM SIGMOD International Conference on Management of Data, pp. 49–60, 1999.

[6] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp.226–231, 1996.

[7] K.G.Sharma , A.Ram and Y.P.Singh, “Efficient Denity Based Outlier Handling Technique in Data Mining”, Proc. 1st international Conference on Computer science and Information Technology, CCSIT,Part 1, pp. 542-550, 2011.

[8] Vishal Bhatt, K. G. Sharma, Anant Ram, “An Enhanced Approach for LOF in Data Mining”, Proc International Conference on Green High Performance Computing, pp. 1–3, 2013.

[9] A.Chiu, and A.Fu, “Enhancements on Local Outlier Detection”, Proc.the seventh International Database Engineering and ApplicationsSymposium (IDEAS’03), pp. 298-307, 2003.

[10] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.