IJITKM **Volume 6 • Number 2 • July-December 2013 pp. 170-174**

A Comparative Analysis of Outlier Mining Techniques with Emphasis on Density Based Technique: Local Outlier Factor

Aditya Dixit [1], Dr. Deepak Dembla [2], Sanjay Tiwari[3]
[1]M. Tech. Research Scholar, Department of Computer Science & Engineering, AIET, Jaipur, INDIA
[2] Professor, Department of Computer Science & Engineering AIET,Jaipur, INDIA
[3]Assoc Prof, Department of information technology, AIET, Jaipur, INDIA
matforcat@gmail.com, sanjay76_tiwari@yahoo.com, deepak_dembla@yahoo.com

Abstract- Outlier can be termed as an observation that does not follow the normal characteristics of the system, hence it is very far from other data points in the system. This property of outlier makes it very suspicions that it may be generated by a different mechanism followed by other data points. The identification of outliers in the area of data mining can lead to Very useful and meaningful knowledge and can be used in practical applications areas such as public safety, public health, climatology, and financial services. In many areas of KDD applications such as frauds in E-commerce the identification and analysis of these abnormal data points is more important than normal data points. There are different approaches for outlier detection; one of them is density based approach. In this paper we provide a detailed comparison of density based outlier finding algorithms such as LOF and its variants LOF' and LOF". Section II discusses the density based approach and LOF. Section III discusses the improvement on LOF, LOF' and LOF". A Complexity comparison is given section IV. Finally, Section V concludes with a summary of those outlier detection algorithms.

Keywords- Outlier, local reachability density, density based approach.

I. INTRODUCTION

Data mining is a collection of steps of identification of previous unknown and useful extracting valid, information pattern in large datasets which can be used in organization decision making [1]. This is also Known ad (Knowledge Discovery in Databases) KDD. However, this task is very complicated as there are a lot of issues such as data redundancy, unavailability of attributes values, and outlier [2].

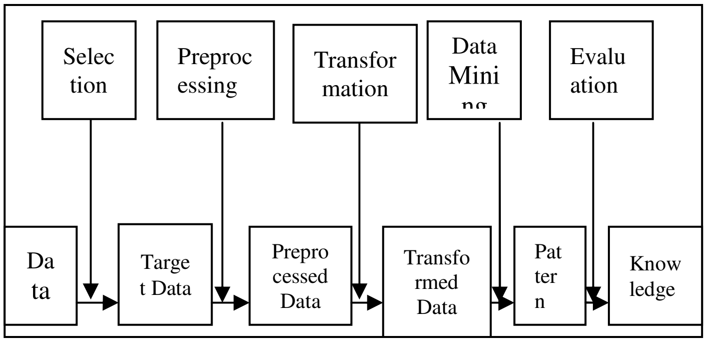

Figure1: KDD Process

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" [2]. From this definition, we can say that Outliers are deviant cases.

Existing work in outlier detection regards being an outlier as a binary property. Outlier detection is an important challenge in data mining with a number of applications from data mining perspective, Recently, many approaches have been proposed for discovering outliers, which can be organized into five categories: distribution oriented, distance oriented, depth oriented, density oriented and clustering oriented [3][4][5][6][7].

There were, a few analysis has been conducted on outlier data identification for large datasets. Distribution oriented methods was initially introduced by the statistics community. In these methods, the data points are mapped to a model with the help of a stochastic distribution, and data points are identified to be outliers on the basis of their relationship with this model. However, when the no of dimensions are increased, it becomes more difficult to apply and inappropriate to estimate the multidimensional distributions of the data points [6].

Distance oriented approach was originally proposed by Knorr and Ang [8]. Further, Ramaswamy et al. [9], had modified distance oriented outlier detection algorithm: the top n points with the maximum D_k are considered outliers, where $D_k(p)$ denotes the distance of the k-th nearest neighbor of p. it is being done with the help of a cluster algorithm to partition a dataset into various groups. Pruning and batch processing on these groups could increase efficiency for outlier detection [10]. Deviation oriented outlier detection does not based on statistical tests or distance oriented measures to identify exceptional data points. Instead, it identifies outliers by examining the main features of data points in a group. Data Points that "do not follow" this description are considered outliers. Hence, in this approach the term deviations are typically used to mane to outliers [1][11].

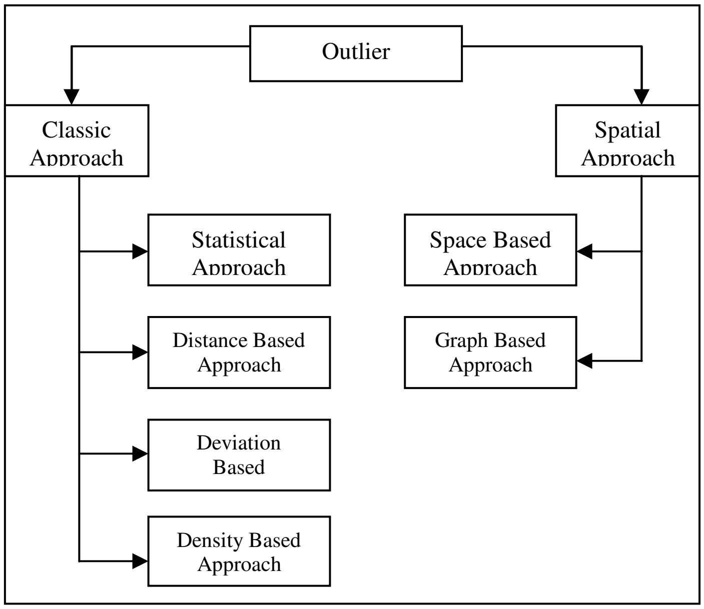

Figure2: Classification of Outlier Detection Techniques

II. RELATED WORK

Density oriented approach was proposed by Breunig et al. [12]. It is based on the local outlier factor (LOF) of each point, which depends on the local density of its neighborhood. Clustering oriented outlier detection techniques considered small clusters as outliers [12] or identified outliers by excluding clusters from the original dataset [12].

III RESEARCH METHODOLOGY

Density oriented approach is based on the concept of neighborhood of the data point. On the basis of neighborhood the density is calculated and there are two main region identified such as low-density region and high-density region. The outlier is the data points found in the low density regions [12]. To estimate distribution density distribution for the given data points and identifying those as outliers which are residing in low-density regions is the basis of this technique. Statistical and distance based outlier detection both depend on the overall or global, distribution of the given data sets of data points, D. However, data are usually not uniformly distributed [12].

- **DENSITY BASED APPROACH**

Density-Based approaches compute the density of regions in the data and find the data points in low dense regions as outliers. The algorithm assigns an outlier score to any given data point, depending on its distance from its local neighborhood. This brings us to the concept of local outliers. An object is a local outlier if it is outlying relative to its local neighborhood, more appropriately with respect to the density of the neighborhood. The algorithms follow this approach are LOF, LOF', LOF".

1. LOCAL OUTLIER FACTOR (LOF):

In [12], Breunig proposes density-based approach which defines the local outlier factor (LOF) for an object in dataset. It is local in that the degree depends on how isolated the object is with respect to the surrounding neighborhood. They give a detailed formal analysis showing that LOF enjoys many desirable properties. In this method a local outlier factor (LOF) for each data point in the dataset, indicating its degree of outlierness is calculated.

Necessary definitions to explain LOF concept are given below. For further details [12] can be referred.

- (k-distance of an object x):-

For any positive integer k, the k-distance of object x, written as k-distance(x), is defined as the distance d(x,o) between x and an object o □ D such that:
(i) for at least k objects o’□D \ {x} | d(x,o’) <= d(x,o), and
(ii) for at most k-1 objects o’□D \ {x}|d(x,o’) < d(x,o).

- (k-distance neighborhood of an object x):-

It is the set of objects o for which distance from x is not greater than the k-distance. Written as $N_k(x)$.

- (reachability distance of an object x w.r.t. object o):-

For the natural integer k. The reachability distance of object y with respect to object x is defined as
reach-distk(x, o) = max { k-distance(o), d(x, o) }.

- (local reachability density of an object x denoted as $lrd_{Minpts}(x)$):-

There are two parameters that define the notion of density:
(i) a parameter MinPts specifying a minimum number of objects;
(ii) a parameter specifying a volume.
The local reachability density of an object x can be calculated dividing one by the average reach ability distance based on the MinPts-distance neighborhood of x.

- (local outlier factor of an object x):-

The LOF of x is defined as

$$LOF_{MinPts(x)} = \frac{\sum_{o \square N_{MinPts(x)}} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(x)}}{|N_{MinPts(x)}|}$$

2. IMPROVEMENT ON LOF

2.1 LOF’[13]:

LOF is a density based algorithm to identify outliers. The local reachability density is used as an indication of density for a data object. We argue that MinPts-tsdist already covers this logic. In view of this, LOF’_ is defined as a simpler formula for ease of understanding, and also simpler computation. This variant of LOF bears more clear meaning and follows similar properties as LOF. The new notion of LOF’ is given below:

- **Local Outlier Factor’**

$$LOF'_{MinPts(x)} = \frac{\sum_{o \square N_{MinPts\text{-}dist(x)}(x)} \frac{MinPts\text{-}dist(x)}{MinPts\text{-}dist(o)}}{|N_{MinPts\text{-}dist(x)}(x)|}$$

LOF’ defined here is the average ratio of MinPts-dist of an object and that of its neighbors within MinPts-dist. A large MinPts-dist means that the density is low since the distance to the nearest MinPts neighbors is large. With this new definition, the components reachability distance and local reachability density needed in the LOF formula are not required anymore. LOF’ captures the degree of outlierness in a similar way as LOF but provides a clear and simple way of formulation.

2.2LOF”[13]:

In many cases outlying objects may be quite close to each other in the data set, so that they form small groups of objects having outlierness. Since MinPts reveals the minimum number of points to be considered as a cluster, if the MinPts is set too low, the groups of outlying objects will be wrongly identified as clusters. On the other hand, MinPts is also used to compute the density of each point, so if MinPts is set too high, some outliers near dense clusters may be misidentified as clustering points. To overcome this problem the neighborhood is divided into [10]:
(1) Neighbors in computing the density and
(2) Neighbors in comparing the densities.
The new notion of LOF” is given below:

Local Outlier Factor”

$$LOF''_{MinPts1,MinPts2} = \frac{\sum_{o \square N_{MinPts1\text{-}dist(p)}} \frac{lrd_{MinPts2}(o)}{lrd_{MinPts2}(x)}}{|N_{MinPts1\text{-}dist(x)}(x)|}$$

One can put a relatively small value as $MinPts_1$ compared with $MinPts_2$ With this simple amendment, LOF" is able to capture local outliers under different general circumstances.

IV.COMPLEXITY ANALYSIS

Mining local outliers by LOF typically requires three loops over the data. A first loop for finding every object's MinPt-dist and MinPts-nearest neighborhoods. Then a second loop to compute the reachability distance and local reachability density for each object. Finally LOF value of all objects in the database is calculated in the third loop. For LOF' computation, the second loop described above is eliminated since the reachability distance and local reachability density are not needed. As a consequence, only two loops over the data is needed. Also note that the second loop that is saved is more complex than the first loop, since for each data point, it requires the collection of information for the neighborhood of a data point.

The algorithm for finding LOF" is nearly the same as that of LOF except for the first loop. No extra loop is required since the $MinPts_2$-distance neighborhood can be obtained directly from $MinPts_1$-distance neighborhood. For LOF, LOF' and LOF", suppose that objects in a database D of size N is being examined, totally there are N MinPts-nearest neighbors queries in the first loop. The complexity ranges from O(N logN) to O($N^{2)}$ depending on the use of indexing structure and dimensionality of data[13]. The following table shows the complexity analysis of these algorithms:

Algorithm	Complexity	Numerical Data	High Dimensional Data	Categorical Data	Type
LOF	O(n log n)	YES	YES	NO	Density
LOF'	O(n log n)	YES	YES	NO	Density
LOF	O(n log n)	YES	YES	NO	Density

Table1: Complexity Analysis of LOF, LOF', and LOF"

V. CONCLUSION & FUTURE SCOPE

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we reviews related work in outlier detection. In particular, special emphasis on density based outlier detection technique is because, the main limitation of distance based approach is that uses the global view concept of dataset which makes it of very limited usage, as many real-world data sets has a very complex representation structure, where objects are required to be considered for outlierness with respect to their local neighborhoods, not with respect to global data distribution. The basic assumption of density based outlier detection approaches is that the density around an inlier is similar to the density around its neighbors; in the case of an outlier object the density around it is much different from the density around its neighbors. Next, we discuss and compare algorithms of outlier detection based on density with the emphasis is on local outlier factor and its variants.

REFERENCES

[1] Han, J. and Kamber, M., Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2001.

[2] D. Hawkins, "Identification of Outliers", Chapman and Hall, London,1980.

[3] J. Xi, "Outlier Detection Algorithms in Data Mining", Proc. Second International Symposium on Intelligent Information Technology Application, vol.1, pp. 94-97, 2008.

[4] K. Singh and Dr. S. Upadhyaya, "Outlier Detection: Applications and Techniques", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey", ACM Computing Surveys, Vol. 41, No. 3, Article 15, 2009.

[6] Z. Bakar, R. Mohemad, A. Ahmad and M. Deris, "A Comparative Study for Outlier Detection Techniques in Data Mining", Proc. IEEE conference on cybernetics and Intelligent Systems, pp. 1-6, 2006.

[7] V. Barnett and T. Lewis, "Outliers in Statistical Data", 3rd edn. John Wiley & Sons, 1994.

[8] Knorr, E.M., Ng, R. T., Tucakov, V., "Distance-based outliers:algorithms and applications", The VLDB

Journal, 2000, vol. 8, pp.237–253.

[9] S. Ramaswamy, R. Rastogi, and S. Kyuseok, “Efficient algorithms for mining outliers from large data sets”. In Proc. of the ACM SIGMOD International Conference on Management of Data, 2000, pp. 93-104.

[10] Aggarwal, C. C., Yu, S. P., “Outlier detection for high dimensional data”, SIGMOD’01, 2001, pp. 37-46.

[11] M.F. Jiang, S.s. Tseng, C. M. Su., ”Two-phase clustering process for outlier detection. pattern recognition letters”, 2001, vol. 22(6-7), pp.691–700.

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

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