Publication of Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th March 2011. Vol. 25 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW REACHABILITY BASED ALGORITHM FOR OUTLIER DETECTION IN MULTIDIMENSIONAL DATASET

[1]K.SUBRAMANIAN [2]DR.E.RAMARAJ
[1]Lecturer, J.J Arts and Science College, Pudukkottai, Tamilnadu.
[2]Technology Advisor, Madurai Kamaraj University, Madurai, Tamilnadu. Email:
Email : subjjcit@gmail.com, eramaraj62@gmail.com

ABSTRACT

The quality of data is major role to detect novel results from the large voluminous databases. So the outlier detection is important process in KDD. It is another important area of data mining research. Maximum of the outlier are due to human errors. Many types of outlier detection algorithms are dealt with in the literature. This work also proposes a new algorithm for detection outliers. It uses new reachability based method for proposed algorithm. The efficiency of the proposed algorithm is proved by using core histogram, letter, segmentation, pima, breast datasets.

Keywords: *Data Mining Quality Of Data Outlier Detection, Reachability, Multidimensional Dataset*

1. INTRODUCTION

The large amounts of data are collected and stored in databases, increasing the need for efficient and effective analysis methods to make use of the information contained implicitly in the data. *Knowledge discovery in databases* (KDD) has been defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable knowledge from the data. Most studies in KDD focus on finding patterns applicable to a considerable portion of objects in a dataset. However, for applications such as detecting criminal activities of various kinds (e.g. in electronic commerce), rare events, deviations from the majority, or exceptional cases may be more interesting and useful than the common cases. Finding such exceptions or outliers, however, has not yet received as much attention in the KDD community as some other topics have(e.g. association rules).

Outliers are the rare or typical data objects that do not comply with the general behaviour or model of the data. Applications such as fraud detection, customized marketing, network intrusion detection, weather prediction, pharmaceutical research, and exploration in science databases require the detection of outliers. There are many known algorithms for detecting outliers, but most of them are not fast enough when the underlying probability distribution is unknown, the size of the data set is large and the number of dimensions in the space is high. There are, however, applications that need tools for fast detection of outliers in exactly such situations.

In this paper, a new method for finding outliers in a multidimensional dataset is introduced. A reachability for each object in the dataset, indicating its degree of outlier-ness is also introduced. This, to the best of our knowledge is the first concept of an outlier which quantifies how an outlying object exists. The outlier factor is local in the sense that only a restricted neighborhood of each object is taken into account. Our approach is loosely related to density-based clustering. However, no explicit or implicit notion of clusters are posited.

2. EXISTING ALGORITHMS

Most of the previous studies on outlier detection were conducted in the field of statistics. These studies can be broadly classified into two categories. The first category is *distribution-based*, where a standard distribution is used to fit the data best. Outliers are defined based on the probability distribution. Over one hundred tests of this category, called discordancy tests, have been developed for different scenarios [2]. A key drawback of this category of tests is that most of the distributions used are univariate. There are some tests that are multivariate. But for many KDD applications, the underlying distribution is unknown. Fitting the data with standard distributions is costly and may not produce satisfactory results.

The second category of outlier studies in statistics is *depth-based*. Each data object is represented as a point in a k - d space, and is assigned a depth. With respect to outlier detection, outliers are more likely to be data objects with smaller depths. There are many definitions of depth that have been proposed [3] [4]. In theory, depth-based approaches could work for large values of ‘g’. However, in practice, while there exist efficient algorithms for ‘g’ = 2 or 3 [4], [5], [6], depth-based approaches become inefficient for large datasets for ‘g’ ≥ 4. This is because depth-based approaches rely on the computation of k - d convex hulls which has a lower bound complexity of $\Omega(n^{k/2})$ for n objects. Recently, Knorr and Ng proposed the notion of *distance-based* outliers [7], [8]. Their notion generalizes many notions from the distribution-based approaches and enjoys better computational complexity than the depth-based approaches for larger values of ‘g’. In [9] the notion of distance based outliers is extended by using the distance to the ‘g’-nearest neighbour to rank the outliers. A very efficient algorithm to compute the top n outliers in this ranking is given, but their notion of an outlier is still distance-based. Given the importance of the area, fraud detection has received more attention than the general area of outlier detection. Depending on the specifics of the application domains, elaborate fraud models and fraud detection algorithms have been developed. In contrast to fraud detection, the kinds of outlier detection work discussed so far are more exploratory in nature. Outlier detection may indeed lead to the construction of fraud models.

3. PROPOSED ALGORITHM

In this type of outlier, the reachability of the neighbours of a given instance plays a key role. Furthermore an instance is not explicitly classified as either outlier or non-outlier; instead for each instance, a reachability outlier is computed which will give an indication of how strongly an instance can be considered an outlier. Breuning et al. [1], shown the weakness of the distance based method in identifying certain type of outliers. The following definitions are needed in order to formalize the algorithm to detect reachability-based local outliers:

The *reachability distance of an instance x* for each *object y* is calculated. The following formula calculates reachability of objects with respected to instance. Let *‘g’* be a positive integer number. The reachability distance of an instance *x* with respect to the instance *y* is defined as:

$$reach - dist_g(x,y) = \max\{\,(g - distance(y), d(x,y))\}$$

This paper proposes new reachability-based outlier detection algorithm for multi-dimensional databases. The proposed problem is broken down into sub phases. The first phase is to calculate the reachability of each object. The second phase is to find outlier from the databases.

The following procedure is used to calculate reachability. It requires four inputs - Dataset *D*, lower number of neighbours *lbn*, upper number of neighbours *ubn*, distance *‘g’* which means maximum distance between center to objects. *gdis-neighbors* means neighbour objects in ‘g’ distance. The *lrdobjects* means that lower reachability data objects, *N* carries number of objects and *o* objects. This procedure is able to generate details objects and their mutual reachability.

```
Procedure calculate reachability( )
Input: Dataset D, lower number of neighbours lbn, upper number of neighbours ubn, distance ‘g’
Begin
1. reachability ←NULL
2. for each g in {lbn,..., ubn} {
3. KDNeighbors ← gdis-neighbors(D, g)
4. lrdobject ← reachability(KDNeighbors, g)
5. for each p in KDNeighbors
6. $temp\_reachability \leftarrow sum(\frac{lrdobject[o \in N(p)])}{\frac{(lrdobject[i])}{|N(p)|}})$
7. reachability ← max{reachability, temp_reachability}}
8. return top(reachability)
End
Output: reachability of each objects
```

The second step is main procedure which is used to detect outliers. It needs two inputs- dataset *D* and reachability of objects.

```
Procedure outlier detection
Input : dataset D with reachability of each object
Output : Outlier
Begin
Outlier ← NULL
for each object p in D
    for each neighbour object of p
        find the nearest maximum reachability of object
        if reachability is equal to maximum number of object
            Move to  next obj
        Else
            delete from D // Outlier
        Endif
    End for
End for
End procedure
```

4. EXPERIMENTAL AND COMPARATIVE STUDY

This section investigates the efficiency of the new proposed algorithm, compared with NL, when applied on different data sets to detect outliers. The proposed algorithm generates outputs that are identical to the outputs of NL. The performance of the proposed algorithm is reported in terms of CPU time. In these tests, five data sets which are obtained from the UCI Repository of Machine Learning Databases [5] have been tested. These are core histogram letter, segmentation, Pima, Breast datasets.

Table 1: Datasets

S.No	Datasets	Dimension	No. of data points
1.	Letter	16	20000
2.	Core Histogram	32	68040
3.	Segmentation	19	2310
4.	Pima	8	768
5.	Breast	10	699

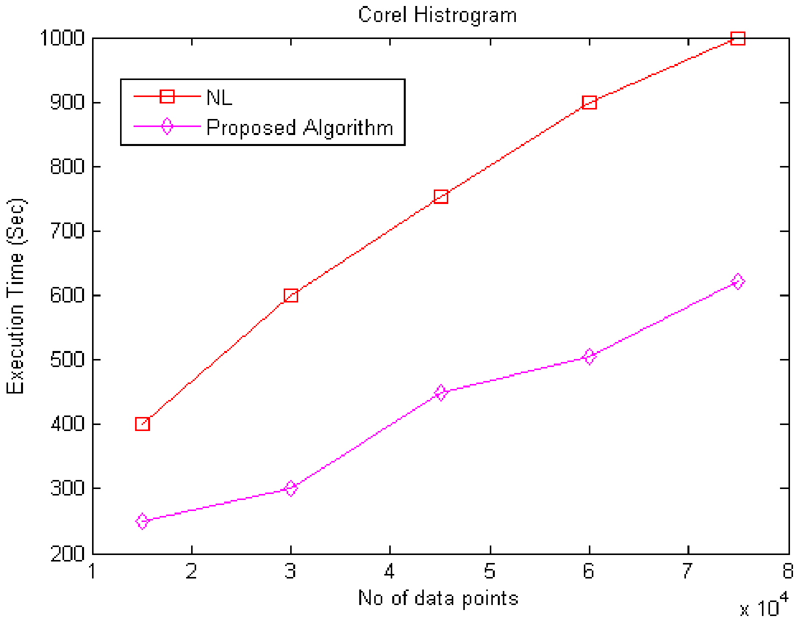

Figure 1 : Comparison of execution time between NL and proposed algorithm using Corel Histogram dataset

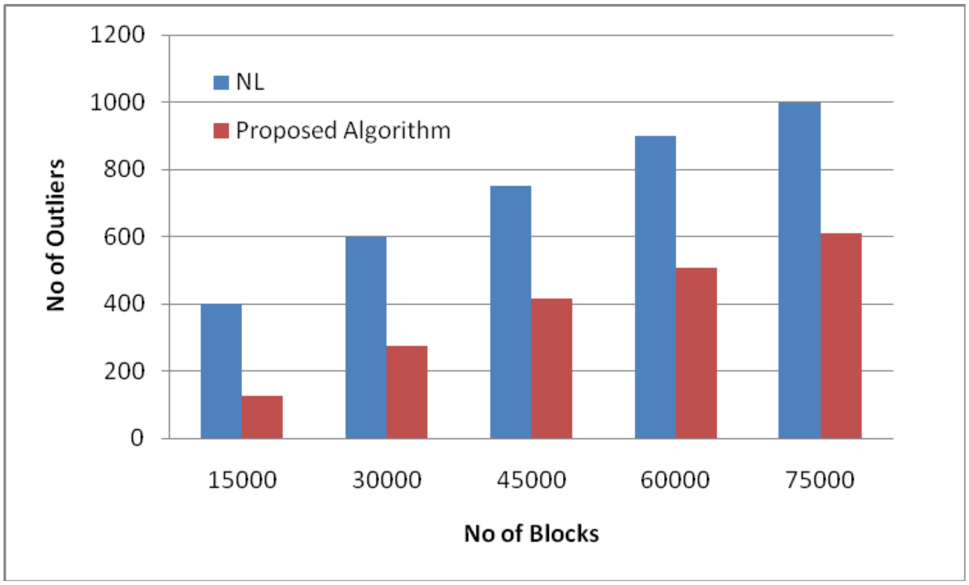

Figure 2 : No of outliers detected in various block size in NL, and proposed algorithm using Corel Histogram dataset

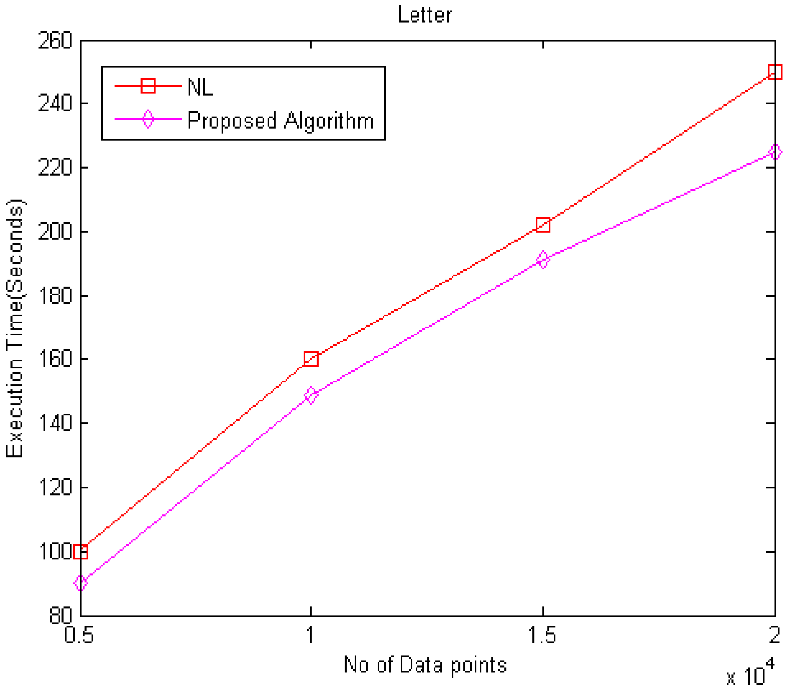

Figure 3 : Comparison of execution time between NL and proposed algorithm using Letter dataset

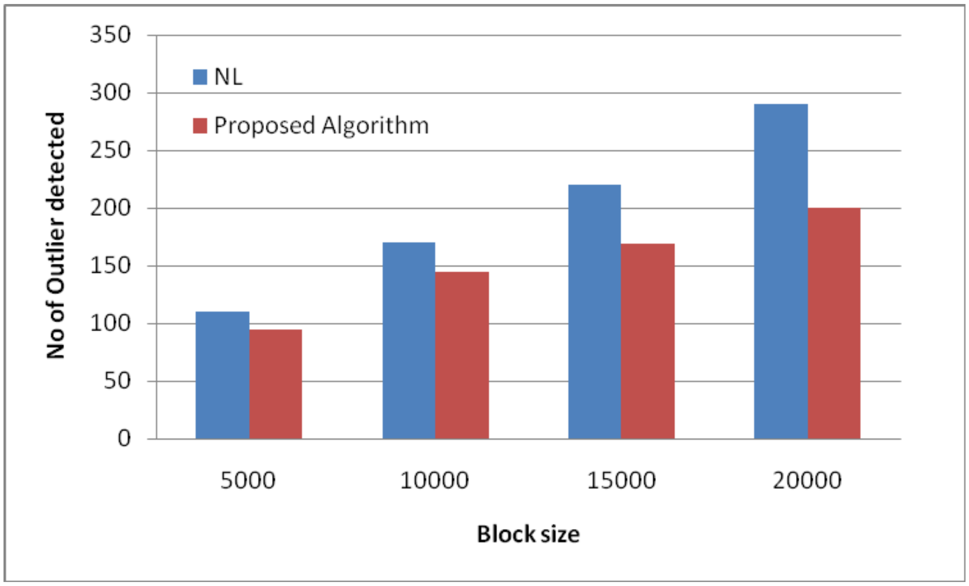

Figure 4 : No of outliers detected in various block size in NL, and proposed algorithm using Letter dataset

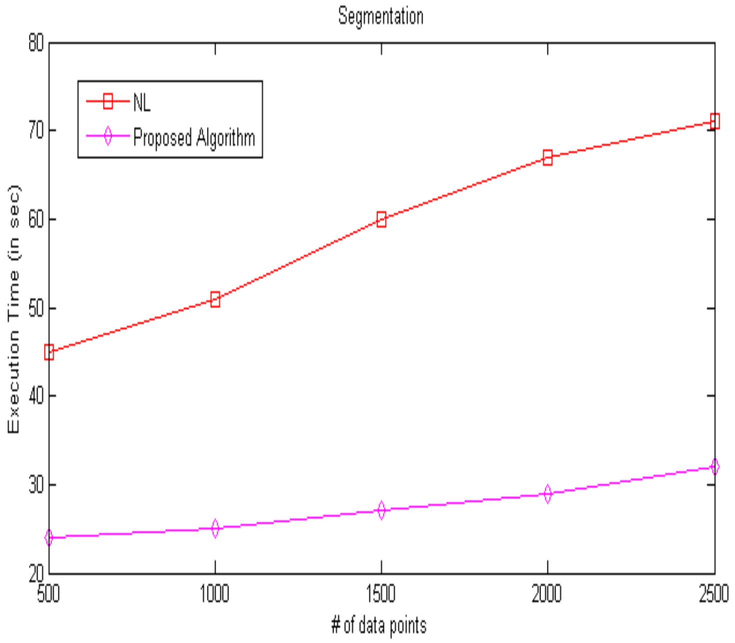

Figure 5 : Comparison of execution time between NL and proposed algorithm using Segmentaion dataset

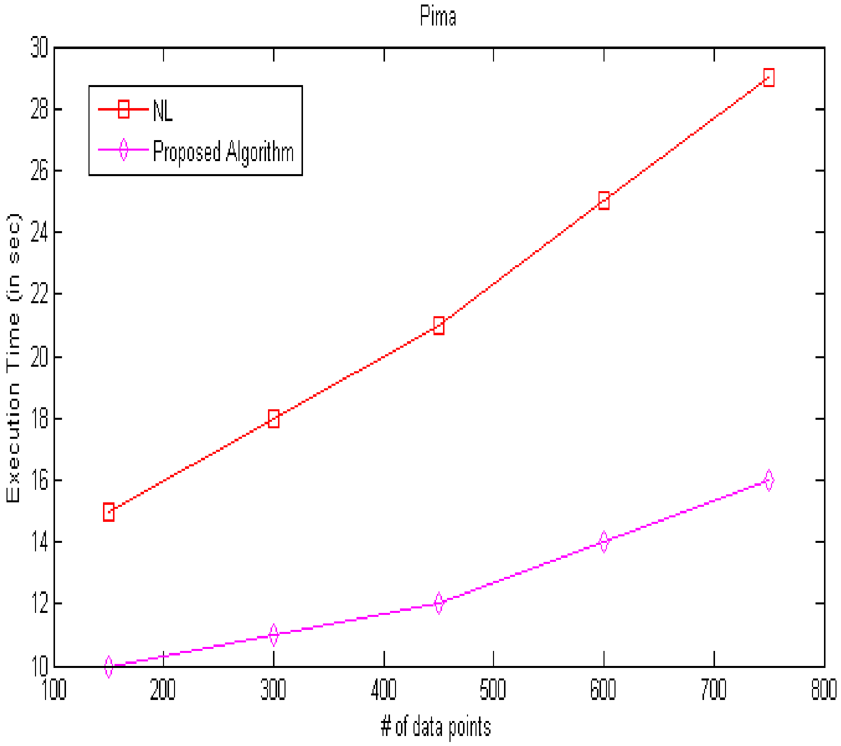

Figure 6 : Comparison of execution time between NL and proposed algorithm using Pima dataset

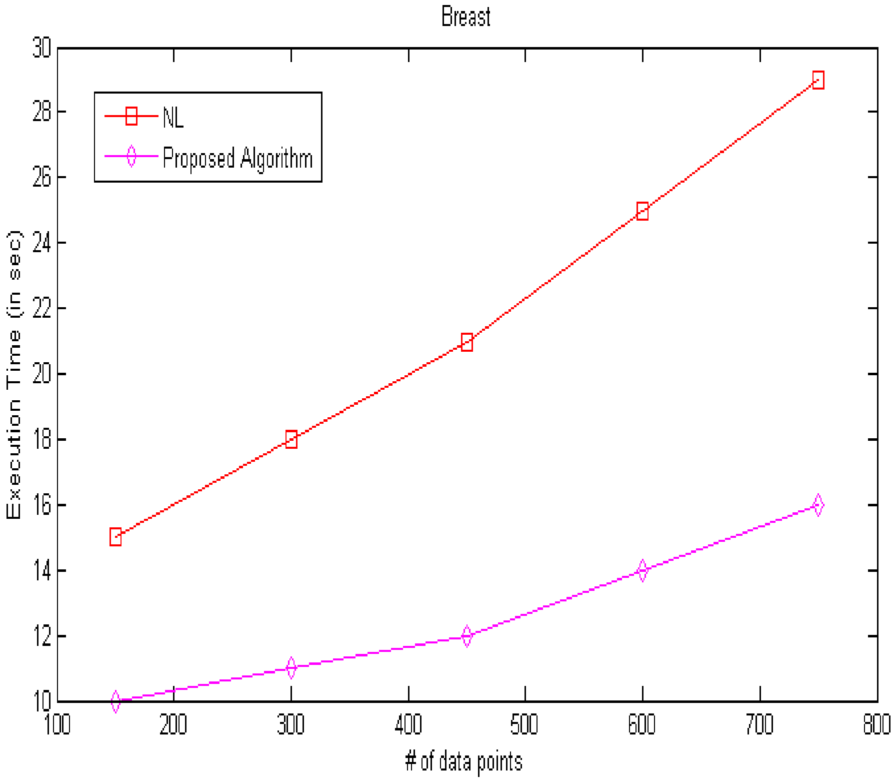

Figure 7 : Comparison of execution time between NL and proposed algorithm using Breast dataset

5. CONCLUSION

The reachability-based outlier detection methods differentiate an object as an outlier on the basis of the reachability between it and its nearest neighbours. This paper proposed a new reachability based outlier detection algorithm. The proposed algorithm has two procedures (i)calculate the reachability of each of the objects (ii) find outliers. This proposed algorithm is compared with distance-based outlier detection algorithm NL. It is important to note that the proposed algorithm performs more comparison operations than NL. The implementation results present a significant increase in efficiency over NL when applied to two synthetic datasets. The present procedure enables us to detect outliers with exactness. It indicates sure cases, no marginal ones.

REFERENCES

[1]. M. Breuning, H. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. *ACM SIGMOD International Conference on Management of Data*, 2000.

[2]. Barnett V., Lewis T.: “*Outliers in statistical data*”, John Wiley, 1994.

[3]. Tukey J. W.: “*Exploratory Data Analysis*”, Addison-Wesley, 1977.

[4]. Preparata F., Shamos M.: “*Computational Geometry: an Introduction*“, Springer, 1988.

[5]. Ruts I., Rousseeuw P.: “Computing Depth Contours of Bivariate Point Clouds, Journal

of Computational Statistics and Data Analysis, 23, 1996, pp. 153-168.

[6]. Johnson T., Kwok I., Ng R.: "*Fast Computation of 2- Dimensional Depth Contours*", Proc. 4th Int. Conf. On Knowledge Discovery and Data Mining, New York, NY, AAAI Press, 1998, pp. 224-228.

[7]. Knorr E. M., Ng R. T.: "*Algorithms for Mining Distance- Based Outliers in Large Datasets*", Proc. 24th Int. Conf. On Very Large Data Bases, New York, NY, 1998, pp. 392-403.

[8]. Knorr E. M., Ng R. T.: "Finding Intensional Knowledge of Distance-based Outliers", Proc. 25th Int. Conf. on Very Large Data Bases, Edinburgh, Scotland, 1999, pp. 211-222.

[9]. Ramaswamy S., Rastogi R., Kyuseok S.: "*Efficient Algorithms for Mining Outliers from Large Data Sets"*, Proc. ACM SIDMOD Int. Conf. on Management of Data, 2000.