Outlier Detection with Two-Stage Area-Descent Method for Linear Regression

HIEU TRUNG HUYNH, MINH-TUAN T. HOANG, NGUYEN H. VO, AND YONGGWAN WON†
Department of Computer Engineering
Chonnam National University
300 Yongbong-dong, Buk-gu, Gwangju 500-757
REPUBLIC OF KOREA

Abstract: — Outlier detection is an important task in many applications; it can lead to the discovery of unexpected, useful or interesting objects in data analysis. Many outlier detection methods are available. However, they are limited by assumptions in distribution or rely on many patterns to detect one outlier. Often, a distribution is not known, or experimental results may not provide enough information about a set of data to be able to determine a certain distribution. Previous work in outlier detection based on area-descent focused on detecting outliers which are solely isolated; it can not detect the outliers clustered together. In this paper, we propose a new approach for outlier detection based on two-stage area-descent of convex-hull polygon. It not only detects outliers clustered together but also shows their location related to the data set. Instead of removing the outlier, this relative location provides a suitable direction for moving the outlier to reduce its effects to linear regression. In addition, this method does not depend on the distribution of data set.

Key-Words: — Outlier detection, convex-hull, polygon, area-descent, linear regression.

1 Introduction

For many applications, finding outliers play an important role in data collection and analysis. It can be more interesting than finding the common patterns. Outliers have been defined informally as data points that are inconsistent with the remainder of data set [1], [2], or observations that deviate so much from other observations as to arouse suspicions that they were created by a different mechanism [3]. Outlier detection has become an important problem in many applications, such as assessment of meter systems, data mining [4], microarray data [5], credit card fraud detection, weather prediction, marketing and customer segmentation, and so on. The outlier detection has also been an important part in the regression.

Many schemes for outlier detection have proposed by researchers [1], [4], [6]-[11]. The early outlier detection methods were based on distribution of the dataset [1]. However, in practice, the distribution is not always known. Knorr E.M. et al proposed a method based on distance [11]. Their method overcame the prior knowledge of distribution, but it can not work well when the data set does not have a uniform density globally. Breunig M.M. et al proposed an approach to measure the strength of an object to be outlier, which is based on the nearest neighborhood for mining local outliers [4]. Brett G. Amidan et al proposed outlier detection using Chebyshev theorem [9]. This method uses the Chebyshev inequality to calculate upper limit and lower limit of an outlier detection value. A method for spatial outlier detection was proposed by Shekhar et al [8], [10]. It is based on the distribution property of difference between an attribute value and the average attribute value of its neighbors. This method had improved by Chang-Tien Luu et al with multiple attributes and multi-iterations that focus on detecting spatial outliers in graph structured data sets [6], [7].

Most previous methods for outlier detection are limited by assumptions of a distribution, which are based on many data points and do not provide a candidate direction to eliminate the outliers. In the regression, shifting outliers can also obtain a better result than removing them. The previous work on outlier detection based on area-descent detected outliers which are solely isolated; it could not detect

† To whom all correspondence should be addressed

outliers that construct a group with a small number of data [12].

In this paper, we present a new approach to detect outliers based on the two-stage area-descent of convex-hull polygon. It does not depend on the distribution of data set and can detect outliers which are in a small group of data.

The rest of this paper is organized as follows. Section 2 reviews related works in outlier detection. In section 3, we present our new approach to identify outliers based on area-descent. The experimental results and analysis are shown in section 4. Finally, we make a conclusion in the section 5.

2 Related works

Most of the early studies on outlier detection are based on statistics [1]. However, in many applications, the distribution is not always obtainable. This limitation is overcome with the distance-based approach proposed by Knorr and Ng. [11].

- *Distance-based outlier detection*

A data point in a data set $\boldsymbol{D}$ is an outlier if its neighborhood contains less than *pct*% of the data set $\boldsymbol{D}$. This approach detects certain kinds of outliers. Because it takes a global view of the dataset, these outliers can be viewed as global outliers. Therefore, it only works well for global uniform density, and can not work well when the subsets of data have different densities. This problem is surmounted by a formal definition of local outliers and a density-based scheme proposed by Breuning et al [4].

- *Density-based local outlier detection*

This method uses the "local outlier factor" *LOF* to measure how strong a point can be an outlier [4]. This scheme can not work well for gradually sparse distribution or density-based clusters that are so closed each other.

- *The Chebyshev outlier detection*

Another method for outlier detection is based on Chebyshev theorem proposed by Brett G. Amidan et al [9]. This method uses the Chebyshev inequality to calculate upper limit (ODV_u) and lower limit (ODV_l) of an outlier detection value. A data with the value that are not within the range of the upper and lower limits would be considered as an outlier. If the distribution of data is non-unimodal, this method may not work well.

Most of former methods do not give any positional information of outlier relative to the dataset from which propose a proper shift direction to reduce its effects. Hence, outliers detected by these methods have to be deleted from the dataset instead of moving them.

- *Area-descent-based outlier detection*

An approach based on area-descent of convex-hull polygon to detect outliers proposed in our previous work [12]. It can detect outliers relying on only two adjacent points on the polygon and shows their location related to the dataset which can infer a proper shift to reduce their effects on linear regression.

The symbols and notations used for reviewing our area descent method are shown in the Table 1.

Table 1: Symbols and Notations

Symbol	Definition
$\boldsymbol{P}$	2D point set
P_i	Data point in **P**
$\boldsymbol{K}$	Convex hull polygon
Q_i	Data point in **K**
S	Area
ΔS	Area descent
Θ	Threshold

Suppose n measurements (attribute values) $y_1, y_2, \ldots, y_n$ $(n \geq 1)$ are made on the referential object $\boldsymbol{x}=\{x_1, x_2, \ldots, x_n\}$. Let $\boldsymbol{P}=\{P_1, P_2, \ldots, P_n\}$ be a 2D point set corresponding to measurements, i.e. $P_i=(x_i, y_i)$.

Firstly, the convex hull polygon is determined; it consists of the most outside points in the data set. Let denote a polygon by $\boldsymbol{K}=\{Q_1, Q_2, \ldots, Q_k\}$ where $\boldsymbol{K} \subset \boldsymbol{P}$, $Q_i \in \boldsymbol{P}$ and $k \ll n$. Let also denote the area of convex hull polygon as S.

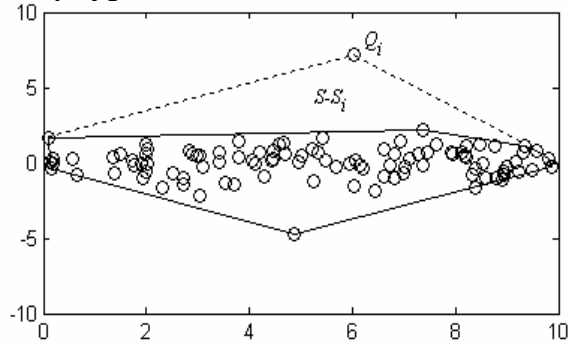


Fig. 1: Detecting outliers by the area descent

For each point Q_i on the polygon $\boldsymbol{K}$, let S_i denote the convex-hull area of point set without Q_i. If the difference between S and S_i is larger than a threshold θ, i.e. $\Delta S_i = S - S_i > \theta$ (*), then Q_i can be viewed as an outlier candidate. There may be many outlier candidates for each polygon $\boldsymbol{K}$ causing many Q_i which satisfy condition (*). Depending on application, all of these candidates will be outliers or only candidates whose area descent is maximum will be outliers, i.e. Q_h is viewed as an outlier if $\Delta S_h = \max_{Q_i \in \boldsymbol{K}} \{\Delta S_i\}$ is larger than the threshold θ. An outline of algorithm for detecting outliers is described as following:

An algorithm for detecting outliers:

1. Detect convex-hull polygon $\boldsymbol{K}$ and its area, S, from point set $\boldsymbol{P}$.
2. For each point $Q_i \in \boldsymbol{K}$, compute $\boldsymbol{P'} = \boldsymbol{P} - \{Q_i\}$, S_i=area of convex-hull polygon of $\boldsymbol{P'}$, and $\Delta S_i = S - S_i$.
3. Compute $\Delta S_h = \max_{Q_i \in \boldsymbol{K}} \{\Delta S_i\}$.
4. If $\Delta S_h > \theta$ then
   - Q_h is an outlier.
   - Remove Q_h from point set $\boldsymbol{P}$, goto step 1.
5. If $\Delta S_h \le \theta$ then stop.

Let we call this algorithm as one-stage algorithm. It can accurately detect outliers that are solely isolated. However, it can not detect outliers that are in even a small group, because their area-descent is too small to pass the predefined threshold. For instance, a simple dataset shown in the Fig. 2 consists of 100 points and two outliers o_1 and o_2 which are very closed each other making a small group.

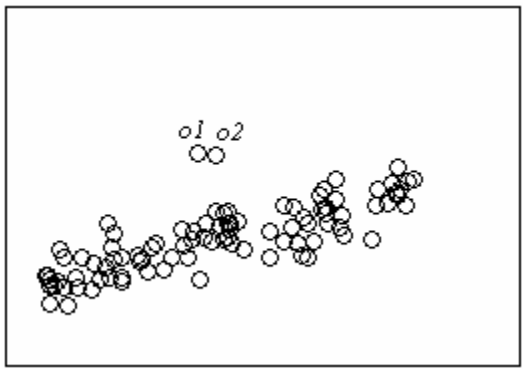


Fig.2: A simple dataset with closed outliers

The value of the area descent corresponding to o_1 will be smaller than that of the other points on the polygon $\boldsymbol{K}$. Hence, o_1 can not be detected as an outlier and thus o_2 can also not be for subsequent iterance.

To overcome this problem, we propose a two-stage area descent algorithm which is detailed in the following section.

3 Two-stage area-descent outlier detection

In this section, we propose a new method to resolve above issues related with outlier detection based on the area-descent. This method bases on two-stage area-descent to detect outliers, it consists of two stages. In the first stage, outliers which are solely isolated are detected using the area-descent algorithm that is the same as the algorithm described in the previous section [12]. The second stage will start when there is no data point that its area-descent is larger than the threshold θ, outliers clustered together will be detected in this stage. In the second stage, we withdraw a subset $\boldsymbol{O}$ consisting of the most outside points of the data set and outliers detected in the second stage should be in this subset. The withdrawing of the subset $\boldsymbol{O}$ consists of two procedures which are repeated: determination of convex hull polygon $\boldsymbol{K}_i$ and remove of the data points involved in constructing $\boldsymbol{K}_i$. $i = 1, 2, \ldots$ The subset $\boldsymbol{O}$ is calculated by $\boldsymbol{O} = \bigcup_i \boldsymbol{K}_i$. These two procedures will stop when the total number of data points in the subset $\boldsymbol{O}$ exceeds a pre-defined threshold value *pts*, let m denote the number of the convex-hull polygon detected in this stage. Thus, polygons K_i, $i=1, 2, \ldots, m$, included in the subset $\boldsymbol{O}$ contain outside points of the dataset and all outliers detected in the second stage should be in these polygons.

Detecting outliers in the second stage is performed from the polygon $\boldsymbol{K}_{m-1}$ to $\boldsymbol{K}_1$. In which, detecting outliers on the polygon $\boldsymbol{K}_i$ based on area-descent is similar to that in the first stage. However, in this stage, the polygons are first determined as previously explained. An algorithm for detecting outliers consisting of two stages is described as following:

The two-stage algorithm for detecting outliers:

Stage 1:

1. Detect convex-hull polygon $\boldsymbol{K}$ and its area, S, from point set $\boldsymbol{P}$.
2. For each point $Q_i \in \boldsymbol{K}$, compute $\boldsymbol{P'}=\boldsymbol{P}-\{Q_i\}$, S_i=area of convex-hull polygon of $\boldsymbol{P'}$, and $\Delta S_i=S-S_i$.
3. Compute $\Delta S_h = \max_{Q_i \in \boldsymbol{K}} \{\Delta S_i\}$.
4. If $\Delta S_h > \theta$ then
   - Q_h is an outlier.
   - Remove Q_h from point set $\boldsymbol{P}$, go to step 1.
5. If $\Delta S_h \le \theta$ go to step 6.

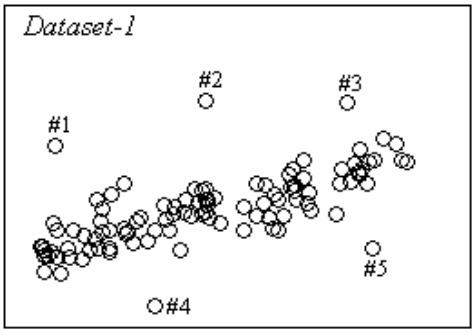
Stage 2:

6. Find a convex-hull polygon $\boldsymbol{K}_1$ from point set $\boldsymbol{P}$.
7. Compute $\boldsymbol{P}=\boldsymbol{P}-\boldsymbol{K}_1$, $\boldsymbol{O}=\boldsymbol{K}_1$.
8. Set $i=2$.
9. Repeat
   a. Find a convex hull polygon $\boldsymbol{K}_i$ from $\boldsymbol{P}$.
   b. Compute $\boldsymbol{P}=\boldsymbol{P}-\boldsymbol{K}_i$.
   c. $\boldsymbol{O}=\boldsymbol{O} \cup \boldsymbol{K}_i$.
   d. $i=i+1$.
10. Until the amount of data points in $\boldsymbol{O}$ larger than pts% of the data set .
11. Set $\boldsymbol{P}_m=\boldsymbol{K}_m$.
12. For $i=m-1$ to 1 do begin
    a. Compute $\boldsymbol{P}_i=\boldsymbol{P}_{i+1} \cup \boldsymbol{K}_i$.
    Repeat
    b. Compute S=area of convex-hull polygon of $\boldsymbol{P}_i$.
    c. For each point $Q_j \in \boldsymbol{K}_i$, compute $\boldsymbol{P'}=\boldsymbol{P}_i-\{Q_j\}$, S_j=area of convex-hull polygon of $\boldsymbol{P'}$, and $\Delta S_j=S-S_j$.
    d. Compute $\Delta S_h = \max_{Q_j \in \boldsymbol{K}_i} \{\Delta S_j\}$.
    e. If $\Delta S_h > \theta$ then
       - Q_h is an outlier.
       - Remove Q_h from point set $\boldsymbol{P}_i$.
    Until $\Delta S_h \le \theta$.
    End for.

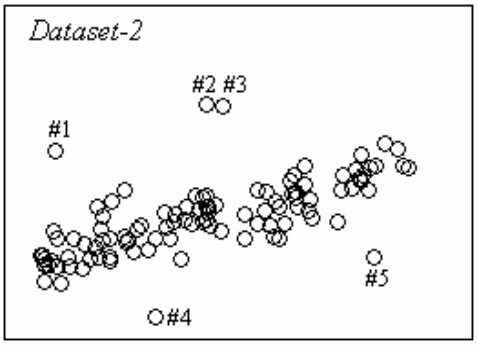
4 Experiments

4.1 Data description

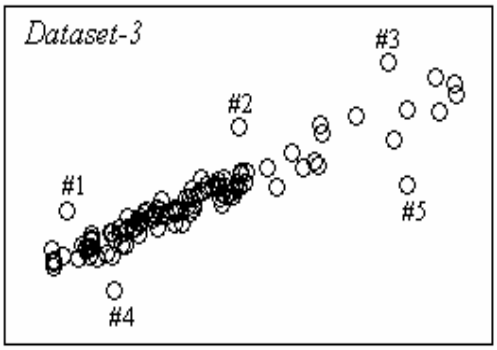
Data sets used in our experiments are similar to data sets in [12], which consist of three data sets with known outliers to evaluate the performance of our method. Each data set contains 100 samples with 5 outliers labeled #1, #2, #3, #4, #5 as shown in the Fig. 3.



(a) outliers are solely isolated



(b) outliers are not isolated



(c) non-uniformly distributed samples (outliers are isolated)

Fig. 3: Three data sets with known outliers

The samples in dataset-1 are uniformly distributed with isolated outliers, the dataset-2 is similar to dataset-1 but outliers #2 and #3 are close to each other making a group of 2. Dataset-3 consists of non-uniformly distributed samples which are similar to difference between some meter system results and primary reference instrument [13].

4.2 Results and Analysis

We evaluate our proposed method by comparing with the distance-based method [11], the density-based

method [4], the Chebyshev Outlier Detection method [9], and our method formerly proposed in [12]. The detailed results are listed in the Table 2. The parameters of the methods are set so that they detect the maximum number of outliers without false positives.

Table 2: Comparative results

Dataset	Dataset-1	Dataset-2	Dataset-3
Area descent (one-stage)	1, 2, 3, 4, 5	1, 4, 5	1, 2, 3, 4, 5
Area descent (two-stage)	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5
Distance-based	1, 2, 3, 4, 5	1, 2, 3, 4, 5	3, 5
Density-based	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 4, 5
Chebyshev	1, 2, 3, 4, 5	1, 2, 3, 4, 5	2, 3, 5

From the results on simulation data, we can see that the isolated outliers in the data set which has uniform distribution are detected accurately by all methods as shown in the column Dataset-1 in the Table 2. In the dataset-2, the outliers #2 and #3 are close to each other, so the former area-descent approach can not detect these outliers, caused by their area-descent is not small enough for detecting. It considers these outliers as another sub-datasets. However, the two-stage approach can detect these outliers accurately.

In the dataset-3, the distribution is non-uniform being similar to the assessment results from biological or medical systems [13]. The distance-based method can only detect outliers #3 and #5. With outliers #1, #2, and #4, their neighborhood contains more than *pct%* of dataset. If we increase the *pct* value, it will detect some inliers as outliers in the sparse region of data set.

The density-based method does not detect outlier #3 in the dataset-3 which lies in the sparse distribution region. Distances between them and their neighbors are quite small compared to distances among their neighbors. The Chebyshev outlier detection method does not detect outliers #1 and #4 in the dataset-3.

As shown in the Table 2, our proposed method can detect accurately all pre-known outliers which are solely isolated or clustered together. The results also show that our method is superior to other outlier detection method for linear regression. In addition, it also shows positional information of outlier relative to the dataset from which propose an appropriate shift direction to reduce its effects on linear regression instead of removing.

5 Conclusion

In this paper, we propose a new approach improved from [12] for the outlier detection based on the area-descent of convex-hull polygons. It can detect outliers clustered together and be much simple that is not dependent upon knowing the distribution of the data but can provide a suitable direction to eliminate effects of outliers on the linear regression.

It is well-known that many algorithms of bioinformatics are great help for clinical performance. These algorithms analyze automatically the collected data and help for the therapy revision as well as for the overall assessment of the patient's behaviors. However, during data collection and analysis there are exist outliers in the dataset. The motivation for our research is to detect and eliminate the outliers to get a better regression to compensate the measurement error of bio-optical signal acquisition system.

Acknowledgements

This work was supported by grant No. RTI-04-03-03 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy (MOCIE) of Korea.

References:

[1] V. Barnett and T. Lewis, "Outliers in Statistical Data", John Wiley, New York, 3rd Edition, 1994.

[2] Ronald K. Pearson, "Outliers in Process Modeling and Identification", IEEE Trans. on Control Systems Technology, Vol.10, Jan. 2002.

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

[4] Breunig M. M., Kriegel H. P., Raymond T. Ng., Sander J., "LOF: Identifying density-based local outliers", Proc. ACM SIGMOD 2000 Int'l Conference on Management of Data, Texas, pp.427-438.

[5] Xuesong Lu, Yanda Li, Xuegong Zhang, "A simple strategy for detecting outlier samples in microarray data", the 8th Int'l Conference on Control, Automation, Robotics and Vision Kunming, December 2004, pp. 1331-1335.

[6] Chang-Tien Lu, Dechang Chen, and Yufeng Kou, "Algorithms for Spatial Outliers Detection", Proc. of the Third IEEE Int'l Conference on Data Mining (ICDM'03), 2003.

[7] Chang-Tien Lu, Dechang Chen, and Yufeng Kou, "Detecting Spatial Outliers with Multiple Attributes", Proc. of the 15th IEEE Int'l Conference on Tools with Artificial Intelligence (ICTAI'03), 2003.

[8] S. Shekhar, Chang-Tien Lu, and P. Zhang, "Detecting Graph-Based Spatial Outlier". Intelligent Data Analysis: An International Journal, Vol. 6, pp.451-468, 2002.

[9] Brett G. Amidan, Thomas A. Ferryman, and Scott K. Cooley, "Data Outlier Detection using the Chebyshev Theorem", Aerospace, 2005 IEEE Conference, March 2005, pp.1-6.

[10] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang, "Detecting Graph-Based Spatial Outliers: Algorithms and Applications(A Summary of Results)", Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and Data Mining, Aug. 2001.

[11] Knorr E. M., Ng R. T., "Finding Intensional Knowledge of Distance-based Outliers", Proc. 25th Int'l Conference on very large Data Bases, Edinburgh, Scotland, pp.211-222, 1999.

[12] Huynh Trung Hieu, and Yonggwan Won, "A Method for Outlier Detection based on Area Descent", Proc. 21th Int'l Conference on Circuit/Systems, Computers and Communications, Vol. 1, pp. 193-196, Jul. 2006.

[13] Richard F. Louie, Zuping Tang, Demetria V.Sutton, Judith H. Lee, and Gerald J. Kost, "Effects of Critical Care Variables, Influence of Reference Instruments, and a Modular Glucose Meter Design", Arch Pathol Lab Med Vol.124, 257-266, 2000.