International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 80-85

Dissimilarity Based Outlier Detection

M. Priya[1], Dr. M. Karthikeyan[2]
[12]*Assistant Professor/Programmer*
[12]*Department of Computer and Information Science, Annamalaiuniversity, India.*

Abstract: *Outlier detection is an important task in data mining with numerous applications. Outliers has been studied in a variety of domains including Big Data, High dimensional data, Uncertain data, Time Series data, Biological data, etc. After studying the commonly used outlier mining methods, this paper presents dissimilarity based outlier detection. Dissimilarity matrix was first formed based on distance between each object of dataset using Gower distance method. Then, Clustering was done using Partitioning AroundMedoids the function PAM method based on the search for k representative objects, and compute total dissimilarity of all objects to their nearest medoid. Experiments were conducted using three dataset such as iris, diabetic and lung cancer datasets. Three clusters were formed for iris dataset, five clusters were formed in diabetic dataset and seven clusters were formed in lung cancer dataset. Finally five outliers in iris, ten outliers in diabetic and twenty three outliers in lung cancer datasets were identified. The experiment results show that to detect the outliers efficiently*

Keywords: *Clusters, Data mining, Dissimilarity, Gower distance, Mediod, Outliers, PAM.*

I. Introduction

Outlier detection is a primary part in the fields of data mining and has attracted the attention of people recently. The task of outlier detection is to identify data objects that are distinctly different from the majority of all objects. An outlier, according to Hawkins, is "an observation that deviates so much from other observations as to arouse that it was generated by a different mechanism".Outlier detection is an important data mining activity with numerous applications, including credit card fraud detection, discovery of criminal activities in electronic commerce, video surveillance, weather prediction, and pharmaceutical researc. Outliers are mainly produced by the following three causes: 1) Data caused by their inherent changes. This change occurs naturally due to data sample, and is uncontrollable. 2) Data result from execute error such as manual operation errors, hacker break and equipment failures. 3) Data that fall into wrong classes. In effective data set, outlier is a small part and recognized as the by-product of clustering. So, outlier is always cancelled or neglected simply. However, researchers gradually realize that certain outlier probably is the real reflection of normal data. So outlier mining becomes an important aspect of data mining. The types of outliers can be classified into three various classes namely, Point outliers which deals with multidimensional data types, Contextual outliers based on the dependency oriented data types such as discrete sequences, time-series, data and graphs. Every instance to a context is defined using the attributes such as Contextual attributes and Behavioural attributes and Collective outliers states that the individual data instance is not an outlier whereas a collection of related data may form an outlier. Avast number of unsupervised, semi supervised and supervised algorithms are found in the literature for outlier detection. Nowadays, the classical technologies of outlier mining can be divided into five categories: statistic-based methods, distance-based methods, cluster-based methods, density-based methods and deviation-based methods.

Above classical methods have respective advantages in application, but they all have some limitation in certain aspects. So, based on dissimilarity, this paper proposes an outlier mining algorithm. The paper is organized as follows. In Section 2, we present the review of related literature and our motivation. In Section 3, the proposed outlier detection algorithm and its related definitions will be described in detail and a performance evaluation is made.In Section 4, the experimental results are analysed. In Section 5 concludes this paper.

II. Literature Review

In this section, we review existing techniques of outlier mining.According to Karanjit Singh and Dr. Shuchita Upadhyaya (2012) has proposed Outlier Detection: Applications and Techniques, in this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details [1]. RashiBansal, Nishant Gaur and Dr.Shailendra Narayan Singh(2016) proposed Outlier Detection: Applications and Techniques in Data Mining,in this review paperwe wish to gain an improved perceptive of various research on outlier detection and analysis for our well-being as well as for those who are the beginners in this field[2]. Statistical-based methods, assuming that data follow a standard

distribution, detect outliers by finding objects that deviate from the distribution (Hubert, Rousseeuw, and Segaert 2015).In distribution-based methods, the observations that deviate from a standard distribution are considered as outliers [5]. Clustering-based methods build clusters model to characterize underlying data behavior. In general, some data objects in clusters that are far away their cluster centers are considered as outliers. For example, cluster histograms Aggarwal (2012) were used for modelling and guiding outlier detection. In Perozzi, Akoglu, Iglesias Sánchez, and Müller (2014), clustering techniques detect outliers in large-attributed-graphs. However, clustering-based approaches must build a clustering model, which limits the outlier-detection performance.Distance-based methods discover outliers by computing distances between an object and other objects in a data set. One of these methods is called local outlier factor (*LOF* Breunig et al., 2000), in which distances are computed by comparing density estimate for each object with its *k* nearest neighbors.Raghav Gupta et al (2016) proposed density based outlier detection technique.A density based approach which used to measure standard deviation method to identify that a data point is an outlier or not [3]. Ana Arribas-Gil and Juan Romo (2013) have proposed shape outlier detection based on the relation between the band depth and the epigraph index. The points which lie below the parabola and the points which are closer to the parabola form a curve with typical shape, whereas the most distant ones form the outliers. From the literature survey, we propose dissimilarity based outlier detection algorithm, in general deal with identifying which objects are potential outliers during the process of clustering. These outliers are eliminated in the final clustering process.

III. Proposed Method

In this section, the proposed algorithm, and its related concept will be introduced in detail.

3.1 Dissimilarity Matrix

Let *D* be a database,an $n \times p$ objects-by-attributes matrix, where rows stand for objects and columns stand for variables.The combination of variables of mixed type into a single dissimilarity matrix slightly extends a definition of Gower (1971), by covering also ordinal and ratio variables. Use distance measure between 0 and 1 for each variable $d_{ij}^{(f)}$.This result in the dissimilarity $d(i,j)$as defined as equation (1)

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \quad \in [0,1] \qquad (1)$$

Where

$d_{ij}^{(f)}$ = contribution of variable f to $d(i,j)$, which depends on its type:

$f$binary or nominal$d_{ij}^{(f)}$ = 0 if $x_{if} = x_{jf}$ and $d_{ij}^{(f)}$ = 1 otherwise,

$f$interval-scaled defined as equation (2)

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}} \qquad (2)$$

$f$ordinal or ratio-scaled: compute ranks r_{if} andtreat these z_{if} asinterval-scaled as defined as equation (3),

$$z_{if} = \frac{r_{if} - 1}{max_h r_{hf} - 1} \qquad (3)$$

and$\delta_{ij}^{(f)}$ = weight of variable f:$\delta_{ij}^{(f)}$ = 0 if x_{if} orx_{jf} is missing,$\delta_{ij}^{(f)}$ = 0 if x_{if}=$x_{jf} = 0$ and variable f is asymmetric binary,$\delta_{ij}^{(f)}$ = 1 otherwise.

A$(n\ x\ n)$ dissimilarity matrix, where$d\ (i,j) = d\ (j,i)$ measures the "difference" or dissimilarity between the objects i and j.$\delta(i,j)$theDistance between i-th and j-th objects. These distances are entries of the dissimilarity matrix (d) asdefined as equation (4),

$$\text{d} = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,n} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,n} \\ \vdots & \vdots & & \vdots \\ \delta_{n,1} & \delta_{n,2} & & \delta_{n,n} \end{pmatrix} \qquad (4)$$

3.2 Partitioning Around Medoids

The function *pam*is based on the search for k representative objects, called medoids, among the objects of the dataset (Kaufman and Rousseeuw 1987). These medoids are computed such that the total dissimilarity of all objects to their nearest medoid is minimal that is the goal is to find a subset $\{m_1, \cdots, m_k\} \subset \{1, \cdots, n\}$which minimizes the objective function as defined as equation (5)

$$\sum_{i=1}^{n} \min_{t=1,\cdots,k} d(i, m_t) \qquad (5)$$

Each object is then assigned to the cluster corresponding to the nearest medoid. That is, object i is put into cluster v_i when medoid m_{vi} is nearer to i than any other medoidm_w, or

$d(i, m_{vi}) \leq d(i, m_w)$for all $w = 1, \cdots, k$ (6)

Compute total dissimilarity distance of all objects to their nearest medoid, it yielding clusters. In that clusters out of the dataset objects remaining objects are outliers. Finally we can detect the outlier by the result of clusters.

3.3 Outlier Detection Algorithm

Generally, outlier keeps away from normal data. Namely, they deviate from the center of data set, and have small quantity. So, the outlier detection focus on finding the data objects which are very dissimilar to the other data objects in some dataset. In our approach, to find out the outliers of each object of data set must be calculated based on dissimilarity matrix. The Outlier Detection Algorithm shown in Table 1.

Table1. Algorithm

Outlier Detection Algorithm

Input:$D = \{d_1, d_2 \cdots d_n\}$ (dataset to be clustered),k value.
Output: $M = \{m_1, m_2 \cdots m_n\}$(clusters medoids), $O = \{o_1, o_2 \cdots o_n\}$(outliers).
Begin
L = {l(d) | d = 1,2,…, n } (set of cluster labels of D)
for each m_i € M do
$m_i \leftarrow e_j$ € D ;(e.g. random selection)
end if
Dissimilarity ← Calculate Dissimilarity Matrix (D, metric);
else
Dissimilarity ← D;
end repeat
for each e_i € D do
l(e_i) ← argminDissimilarity(d_i ,Dissimilarity, M);
end
for each m_i € Mdo
Mtmp ← Select Best Cluster Medoids(D, Dissimilarity, L);
end
outliers = order(Mtmp)
End

3.5 Performance Evaluation

In order to show the effectiveness of the proposed method,performance evaluation based on real-world datasets were conducted. In the experiment, show that to detect the outliers efficiently.

3.6 Metrics for Measurement

Forperformance evaluation of the algorithms, we use two metrics, namely Recall and Precision, to evaluate the detection results.Then the Recall and Precision are defined as follows.

$$\text{Recall}=\frac{number\ \ of\ the\ identified\ \ outliers\ \ by\ algorit\ hm}{number\ \ of\ the\ expected\ \ outliers} \quad (7)$$

$$\text{Precision}=\frac{number\ \ of\ the\ identified\ \ outliers\ \ by\ algorit\ hm}{number\ \ of\ the\ identified\ \ objects\ \ by\ algorit\ hm} \quad (8)$$

The possible maximum value of Recalland Precisionis 1, and the possible minimum value of Recall and Precisionis 0. The bigger the value of Recall and Precision is, the better the results of outlier detection.

IV. Experimental Results

For our experiments, we have used three datasets. The evaluation is performed on a number of *UCI* data sets with different characteristics.We also applied the proposed method to real-worlddatasets.The Iris dataset contains 150 objects that were grouped into three clusters. The Diabetic dataset contains 768 objects that were grouped into six clusters. The Lung Cancer dataset contains 135 objects that were grouped into five clusters. The datasets are obtained from the University of California, Irvine (UCI) machine learning repository. In this paper, we select three clusters as normal objects and select five objects from the remaining cluster as outliers from Iris dataset, five clusters asamong the six clusters normal objects and select ten objects from the

remaining cluster as outliers from Diabetic dataset and seven clusters asamong the five clusters normal objects and select twenty threeobjects from the remaining cluster as outliers from Lung Cancer dataset.

The experimental parameters for Data Analysis of outputs and Comparison of results are depicted in Table 2.

Table 2. Analysis of Outputs and Comparison of Results

Dataset	Number of objects (n)	k	Number of clusters		Outlier Objects Identified by Proposed Method	Actual Outlier Objects
			Identified by proposed method	Actual clusters		
Iris	150	4	3	3	5	5
Diabetic	768	3	6	5	10	12
Lung Cancer	135	4	7	5	23	23

In Fig.1shows thatthe graphical representation of analysis of outputs and comparison of results for clusters and Fig.2shows thatgraphical representation of Analysis of outputs and Comparison of results for outliers.

Fig.1. Comparison results for clusters.

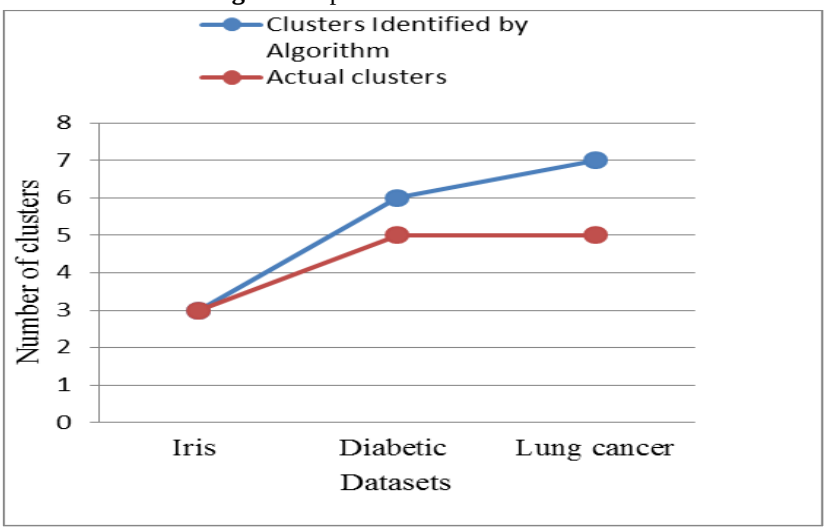
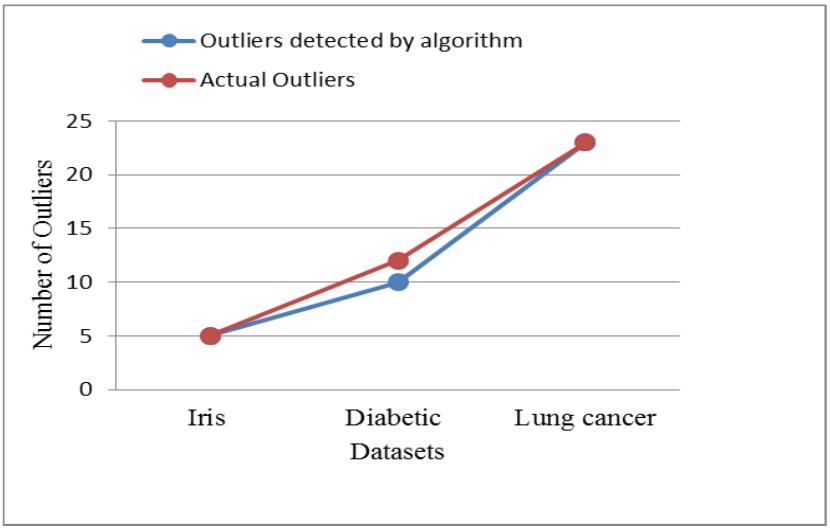


Fig.2.Comparison results for outliers.



After detecting the outliers, we evaluate the performance metric measurements, recallin equation (7) and precision in equation (8) for the three datasets. Table3. shows the performance evaluation recall, clusters and outlier objectsof three datasets. Fig. 3 shows that the graphical representation of measurement metrics.

Table3. Measurement metrics

Dataset	Clusters	Outlier objects	Recall
Iris	3	5	1
Diabetic	6	10	0.83
Lung cancer	7	23	1

The precision of proposed outlier detection algorithm is same as recall ratio for three datasets. The clustering performance improves as more and more outliers or weakly relevant objects are removed. It is evident that the number of outliers increases with the number of objects. Experimental set up shows thatto detect the outliers efficiently.

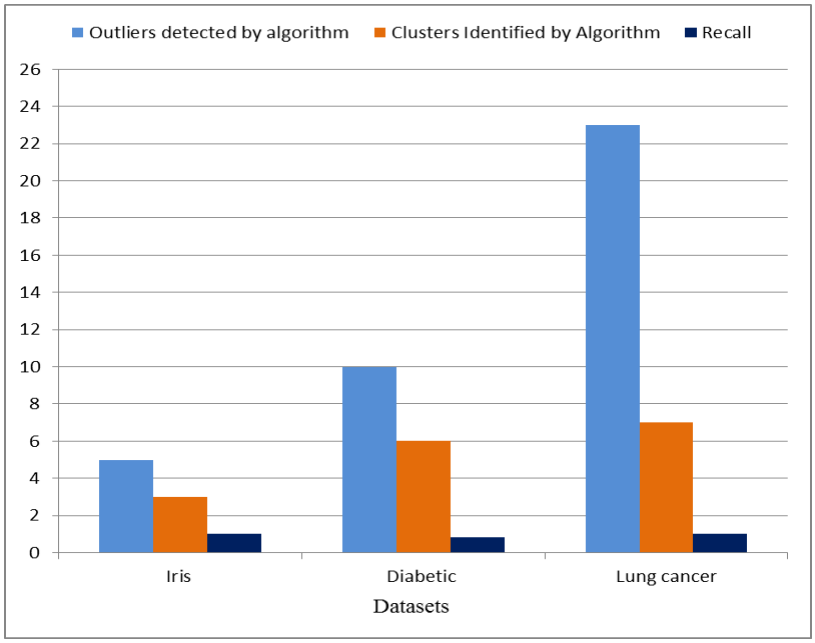


Fig. 3. Recall, clusters and outlier objects of three datasets

V. Conclusion

The purpose of outlier mining is to find small groups of data that are exceptional within a large amount of data. Mining of such outliers is important for many applications. This outlier may be due to the unavailability or distortions in the data collection stage that consists of irrelevant or weakly relevant data objects. From the algorithm, it is shown that by choosing a valid outlier objects, the overall performance of the algorithm can be improved.This paper presents dissimilarity based outlier detectionin data mining. Experimental results show that this algorithm has better recall and precision. So it is more suitable for massive data.

References

Journal Papers

[1]. Karanjit Singh and Dr. ShuchitaUpadhyaya(2012),Outlier Detection: Applications and Techniques, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.
[2]. Rashi Bansal, Nishant Gaur, Dr. Shailendra Narayan Singh(2016), Outlier Detection: Applications and Techniques in Data Mining, IEEE 2016, 6th International Conference - Cloud System and Big Data Engineering, (Confluence) 978-1-4673-8203-8/16.
[3]. Raghav Gupta and Kavita Pandey (2016),Density Based Outlier Detection Technique, Springer Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing.
[4]. Knorr E, Ng R, Finding Intentional Knowledge of Distance-Based Outliers,Proc. of the VLDB Conf. Edinburgh: Morgan KaufmannPublishers, 1999,pp.211-222.
[5]. Arning A, Agrawal R, Raghavan, A Linear Method for Deviation Detection in Large Database, Proc. of the KDD Conf., Portland: AAAI Press,1996,pp.164-169.
[6]. L. Duan , et al. , Cluster-based outlier detection, Ann. Oper. Res. 168 (1) (2009) 151–168 .
[7]. J.-k. Min,An efficient outlier detection algorithms based on data clustering over massive data, Database Res. 31(3)(2015) 59–71.
[8]. D. Yu , G. Sheikholeslami , A. Zhang , Findout: finding outliers in very large datasets, Knowl. Inf. Syst. 4 (4) (2002) 387–412 .
[9]. Tu Lihong, Tong Haiyan, Yang Liping, Research of Outlier Mining Based on Dissimilarity,Computer and Digital Engineering,2008,36(1), pp.16-19
[10]. Breunig, M. M. , Kriegel, H.-P. , Ng, R. T. , & Sander, J. (2000),Lof: identifying density-based local outliers,In ACM SIGMOD record: 29 (pp. 93–104). ACM .

[11]. Hubert M, Rousseeuw, P. J. & Segaert P. (2015),Multivariate functional outlier detection, Statistical Methods and Applications, 24 (2), 1–26 .
[12]. Perozzi, B.,Akoglu, L.,Iglesias Sánchez, P.,&Müller, E. (2014),Focused cluster- ing and outlier detection in large attributed graphs,In Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1346–1355), ACM.
[13]. Aggarwal, C. C. (2012), A segment-based framework for modeling and mining data streams,Knowledge and information systems, 30 (1), 1–29.
[14]. Nattorn Buthong, Arthorn Luangsodsai, Krung Sinapirom saran, Outlier Detection Score Based on Ordered Distance Difference, IEEE International Computer Science and Engineering Conference, 2013.

Books:

[15]. Barnet V, Lewis T, Outliers in Statistical Data, New York, John Woley& Sons, 1994.
[16]. Hawkins D, Identification of Outlier, London: Chapman and Hall, 1980.
[17]. J. Han, M. Kamber, J. Pei,Data Mining: Concepts and Techniques: Concepts and Techniques, Elsevier, 2011.
[18]. T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to Data Mining, Library of Congress, 2006.
[19]. Mao Junguo, Duan lijuan, Wang Shi, Shi Yun,Data Mining Principle and Algorithm, Tsinghua University Press, 2007.
[20]. Gower, J. C. (1971), A general coefficient of similarity and some of its properties,Biometrics, 27, 623--637.
[21]. Kaufman, L. and Rousseeuw, P.J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York.
[22]. Blake C, Merz C.UCI Machine Learning Repository, www.ics.uci.edu/~mlearn/MLRepository.html.