OUTLIERS COMPARISON OF VARIOUS KINDS OF CLUSTERS USING THE SAMPLE SURVEYS DATA OF INDIA

[1]Chetan Chadha, [2]Shivang Garg
[1]Student, [2]Student
[1] M.B.A(I.T.), BVIMR, BVDUSDE, New Delhi, India,
[2] B.Tech(C.S.E.), NIT, Uttarakhand, India

Abstract: Data Mining (referred as extracting knowledge from data) is the process of discovering patterns, associations, and links in the huge stack of data which is based on the analysis done through different perspectives. There are many disciplines which are found under data mining some of them are clustering analysis, regression analysis, classification analysis etc. To analyze the data, we need some sort of software which can be used to perform the required analysis on data set. In this paper, we are using WEKA library (.jar file) for data mining and implementing the outlier detection algorithm to detect the outliers in the data set. Outliers are the points in our data set which is distinct from the remaining data set and including of these points may result in inappropriate outcomes. [1] Outliers are the observation that deviates so much from other observation as to arouse suspicion that it was generated by a different mechanism. In this paper, we are studying various data mining techniques to find outliers in our data set and also finding the best suited algorithm for different sample survey data of India (comprises of Inflation rate, Demonetization, UIDAI, and Pan Card) that will help us to find the set of outliers for our data set**.**

IndexTerms **– Outliers, Clustering, Survey, K-mean clustering, Hierarchical clustering, Spectral clustering, Partition clustering, Density-based clustering, Quadrant, Grubb value**

I.INTRODUCTION

Outlier detection is nothing rather than finding the data points whose behavior is very exceptional when compared with the remaining set of data points. Since these outliers have different behavior, it is important to study this extraordinary behavior so that it helps to uncover the valuable knowledge which is hidden behind them. Outliers detection helps the makers to make the profit and to improve the service quality. Thus, finding outliers is an important research with the large number of applications, including the discovery of abnormal activities, weather prediction, fraud detection, marketing, and in finding criminal activities in electronic commerce.

In this paper, for detecting outliers we are making two-dimensional graph whose x-axis and y-axis are denoting the conditions and values on the basis of which data has been classified using clustering algorithm. After this, we had made quadrants for this two-dimensional graph. Then, plot all the data points on the above two-dimensional graph, all the points which do not lie in these quadrants will be referred as outliers.

Data mining is a technique to analyze and retrieve knowledge from a large amount of database and transform it into useful information for future use [2]. Data mining is used in classification, clustering, regression, association rule discovery, sequential pattern discovery, outliers detection, etc. [4]. Data mining is the process of uncovering previously undetected relationships among data items. Data mining is not the single process, it is the multi-stage process. Data mining is the process, in which data is mined by going through different phases which are given as Data Selection, Data Preprocessing, Data Transformation, Data Mining and Interpretation & Evaluation. Data can be mined through two approaches which are supervised and unsupervised learning.

Supervised learning is the approach of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each data sample consisting of an input object and the desired output value. A supervised learning algorithm inputs the training data and produces an inferred function, which will be further used for mapping new examples which are unseen instances. Unsupervised learning is the process of finding hidden structure in unlabeled data. In Unsupervised learning, each data sample consisting of an input object only (no class label). Clustering is an example of a unsupervised learning. It is the task of grouping a set of objects in such a way that objects in the cluster are more similar to each other than to those in other clusters [5].

In this paper, we try to discover the unexpected results that are hidden and must be disclosed for better results in future. Let us take the example of data comprises of the salary of the different age group of individuals, then classification can only be used to classify that data based on our conditions. Let's say individual comes in the range of middle class when his salary within the range of 15k-30k (INR), in range of poor class when his salary within the range of 5k-15k (INR), and in range of rich class when his salary is above 50k (INR). Then, classification helps to list out data with this range based on the age of individual (18-60) years. For this example, let

suppose two- dimensional graph have x-axis as salary and y-axis as age. The range of x-axis as 0k-100k (INR) and for y-axis as 18-60 years. The midpoint for x-axis is 50k (INR) and for y-axis is 39 years. So, the quadrants for this graph are one with age 18-39 years and salary 0k-50k (INR), second with age 18-39 years and salary 50k-100k (INR), third with age 39-60 years and salary 0k-50k (INR) and fourth with age 39-60 years and salary 50k-100k (INR). Then the points which are out of these quadrants will be classified as outliers. Classification will not tell about the data including (individual with age more than 60 years and have salary of more than 50k (INR), individual within age of 18-60 years but not earning any salary), which are nothing but outlier data points and these data points are also important while implementing any policy as these minority factors cannot be left alone.

The rest of paper is organized into five sections. Section I explains the related work. Section II describes the clustering algorithms compared. Section III explains the methodology used in this paper. Section IV presents the experimental result in tabular and graphical forms. Finally, section V concludes the paper.

II. RELATED WORK

Some past researchers have implemented new algorithms while others have improved the data clustering algorithms. Also, some of them have analyzed and compared the existing algorithms related to clustering for data mining. [6] applied various indices to evaluate the performance of various clustering algorithms, these indices are number of clusters formed, numbers of outliers, a time taken, number of data points per cluster, non-clustered instances, a squared error within the cluster and many more.

Some researchers [7] have discussed the technique and detailed the performance of different algorithms based on speed and time. While others [8] present the normal behaviors by using a Gaussian mixture. Bay and Schwabacher [9] give an algorithm which detects outliers nearly in linear time based on distance method. Clustering based techniques for outlier detection regarded small cluster as outliers [10] or helps in identifying outliers by removing clusters from the actual dataset. [11] discussed new technique, which is based on observing the density distribution. [12] discussed an outlier mining method, which is based on a hypergraph model helps to detect outliers in the categorical dataset. [13],[14] used replicator neural network (RNN) to detect outliers. Liu and Jezek [15] have proposed the method for outliers detection in an irregularly distributed spatial dataset .M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander proposed the method for Local Outlier Factor. In this method, an outlier is detected by measuring the local deviation of a given object with respect to its neighbors. This factor is based on local density concept. The k-nearest neighbors are used to compose the object's neighbors. Arning et a1. Proposed a deviation-based method that inspects outliers from the main characteristics of objects and objects in the dataset that deviates from these features are considered as outliers [16].

III. CLUSTERING ALGORITHM

Clustering is a prominent task in mining data sets, which groups related data objects into a cluster. There is the number of clustering algorithms for mining the data. Data mining is nothing but extracting the useful patterns from data set through which we can extract fruitful outcomes. The term clustering is used by a lot of research communities to describe the method of unlabeled data grouping. Clustering makes the group of the data and hence improve the efficiency of the result. The clustering algorithm groups object into subclasses. There are different algorithms whose suitability depends on the types of application. Some of them are K-means Clustering algorithm, a Hierarchical clustering algorithm, a Spectral clustering algorithm, Partition clustering algorithm, Density-based clustering algorithm and Grid-based algorithm.

3.1 Partition Based Clustering

Partition based clustering is the easiest and simplest way of partitioning the dataset. It organized the data elements into no of clusters. Some of the partition-based clustering methods are K-Means, PAM, K-Medoids, CLARA, etc. In this algorithm, the no of clusters and dataset act as the input and output is the set of the cluster. It iterates the data elements for the better clustering result.

3.1.1 K-Means Clustering

K-Means is an iterative clustering algorithm [17], [8] in which samples are classified among specified no of clusters until the desired state achieved. This algorithm used squared error to calculate deviation at each iteration until it reaches the value which can be permissible. The cluster mean of Ki={ci1, ci2, ... , cin}is defined as,

mi= 1/n(summation from 1 to m of ci) (3.1)

Steps of K-Means Clustering:

- Choose k objects randomly from the dataset which acts as the initial cluster points.
- Then pick data element from the dataset and assign them to the cluster to which it is closely association (minimum distance) which is calculated based on the distance between data element and centroid of the respective clusters.
- After classifying all the points, compute the new position of the centroid for each cluster which is the mean value of all the data points in the respective cluster.
- Calculate squared error for all the clusters which are calculated as the summation of the square of the distance between data element and centroid of the respective clusters for all the points in the cluster.
- Repeat step 2, 3 and 4 until the points stop moving, i.e. the mean squared error converges.

Mostly in all the cases, the simple K Means clustering algorithm required more time to cluster the data which makes it unsuitable to be used for large datasets.

3.1.2 Density-Based Clustering

In partition and Hierarchical methods, we can only find spherical-shaped clusters which are one of the biggest disadvantages of these algorithms as these algorithms are unable to find Oval and "S" shaped clusters. But, with the help of density-based clustering algorithm, we can find randomly shaped clusters. With this algorithm, we can denote cluster as dense in data set, which is separated from the sparse region. The main ideology behind this algorithm is to find the arbitrarily shaped clusters.

Steps of density-based Clustering:

- Firstly, we compute the ε-neighborhood for all the data points in the dataset, where ε is the minimum number of data points.
- Select the Core Points out of these data points. Core Points are the points whose ε (found in step-1) is equal or more than the specified value.
- For all the data element belongs to Core points, add the density connected data elements to these Core points, until no further density connected data points encountered.
- Repeat step 2 and 3 until no core points left.

DBSCAN (Density-based Spatial Clustering of Application with Noise) is one of the density-based clustering algorithms which handles the random shaped clusters and noise in the cluster using density connectivity [19]. DENCLUE (Density-based clustering) is a distribution-based algorithm, which works effectively on the dataset with high-level noise and has large no of data elements. It works faster than DBSCAN, it contains large no of parameters which makes it highly effective to find the randomly shaped clusters. It can be applied only on medium and small level datasets due to its non-linear time complexity [20].

3.1.3 Hierarchical Clustering

Hierarchical clustering is the clustering analysis method which assigns data elements into a tree-like structure, in which each cluster has a data element. Hierarchical clustering is generally divided into two categories: -

3.1.3.1 Agglomerative Hierarchical Clustering (AHC)

AHC used the bottom-up approach to classify the data set. Initially, each data element act as a different cluster and later these clusters merges into large clusters until all the data elements merge into a single cluster, means at the end of AHC we left with a single cluster which consists of all the data elements. For merging, data elements find the cluster which is closest in distance to it and combines to form the one cluster and then change the centroid of the combined cluster based on different conditions (Complete Linkage clustering, Single Linkage clustering, Average Linkage clustering and many more). AGNES (agglomerative nesting) is an agglomerative hierarchical clustering method.

3.1.3.2 Divisive Hierarchical Clustering (DHC)

DHC used the top-down approach to classify the data set. DHC algorithm is just opposite of AHC. In DHC initially, all the data elements act as a single cluster and later this single cluster breaks into different clusters until each data element separate to form a single cluster means at the end of DHC we left with the no of cluster equals to the no of data elements in our dataset. DIANA (Divisive Analysis) are examples of divisive clustering method.

IV. RESEARCH METHODOLOGY

To compare different clustering algorithm and to find out the best one for detecting outliers, we had taken dataset from sample survey data of India (comprises of Inflation rate, Demonetization, UIDAI, and Pan Card details). We had applied K-Means clustering as partition based clustering algorithm, DBSCAN as density-based clustering and Agglomerative hierarchical clustering as Hierarchical clustering algorithm for classification. For clustering our dataset, we are using WEKA library (.jar file) which took as input our dataset and give output as the set of clusters. Then, we had converted the output data in the form of JSONObject (key-value Pair). We input JSONObject data to our own OUTLIER_ANALYSIS algorithm which process the JSONObject data and output the outliers which we had shown in this paper. For outlier analysis, we are making two-dimensional graph whose x-axis and y-axis are denoting the conditions and values on the basis of which data has been classified using clustering algorithm. After this, we had made quadrants for this two-dimensional graph and after this, we find the midpoint of x-axis and y-axis to find the centroid of the quadrant. Then, plot all the data points on the above two-dimensional graph with data points lie inside the quadrant as well as outliers data points.

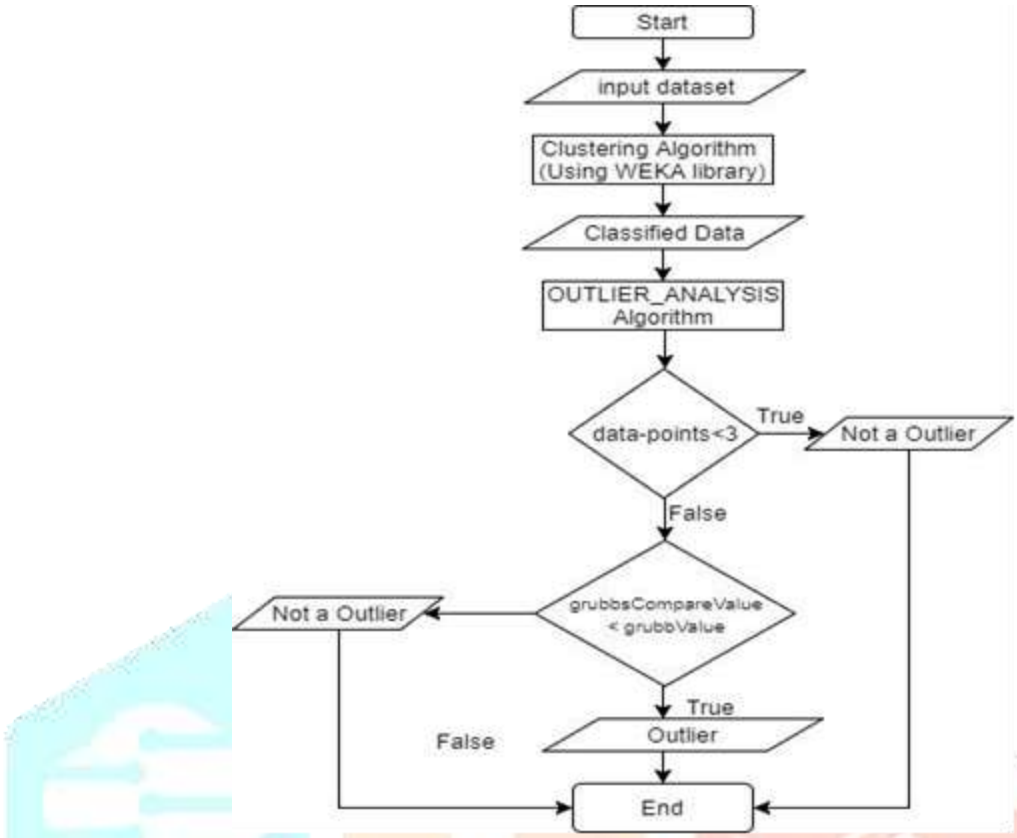

Figure 4.1: Flow graph of our algorithm implemented in this paper

4.1 Outlier Analysis Algorithm

The input (values and Significance level) is described as:

- Values are in the form of JSONObject data, which we get after parsing the data coming from WEKA library (.jar file).
- The significance level is the maximum acceptable risk level for rejecting the null hypothesis in the case where the null hypothesis is true.

Steps of Outer Analysis Algorithm:

- Firstly, find the number of data elements in the clustered data and if it is less than 3 then discard that set.
- If data elements are greater than or equal to 3, then find grubbsValue and grubbsCompareValue for the data elements in clustered data.

$$Gexp=G10=|Xout-X^{-}|sGexp=G10=|Xout-X^{-}|s \qquad (4.1)$$

The suspected outlier is rejected if Gexp is greater than G(α, n).

- To find the grubbsValue, we have to use TDistribution concept in Java, which helps us to find the cumulative probability and inverse cumulative probability of data elements.
- From TDistribution, we can find the critical value of data sets which is further used to find grubbsValue.
- To find grubbsCompareValue we have to find the maximal deviation and standard deviation.
- Maximal deviation acts as a boundary above which data element is classified as the outlier.
- The Maximal deviation is further calculated using the mean value of the data elements and if absolute of mean and data element value is greater than maximal deviation then marked it an outlier.
- After finding both grubbsValue and grubbsCompareValue, we can check if grubbsValue is greater than the grubbsCompareValue, then mark it an outlier. Otherwise, discard the cluster and go with new clustered data until all clustered data finishes.

Since each clustering algorithm gives the different value for input values to this algorithm, it will give the different value of outliers for each algorithm.

4.1.1 Algorithm

Outlier analysis (input values, significanceLevel)

size = Number of data elements in data set
if size < 3
return null
end if
tDistribution = Create a distribution with (size - 2.0) as degree of freedom
criticalValue = tDistribution .inverseCumulativeProbability
((1.0-significanceLevel) /(2.0 * size))
criticalSquareValue = criticalValue * criticalValue
grubbsValue = (size - 1) / (square root of size) * (square root
of(criticalValueSquare / (size - 2.0 + criticalValueSquare)))
standardDeviation = Standard deviation of the data elements in dataset maximalDeviation = 0
for each data element in dataset
if absolute(mean - value) > maximalDeviation
maximalDeviation = absolute(mean - value)
set this data element as outlier
end if
end for
grubbsCompareValue = maximalDeviation / standardDeviation if grubbsValue > grubbsCompareValue
return set of outliers
end if
else
return null
end else

V. RESULTS AND DISCUSSIONS

In this paper, we are calculating the outlier detection accuracy which is based on the accuracy of the clustering algorithm. For that, we have used sample survey data of India (comprises of the Inflation rate, Demonetization, UIDAI, and Pan Card). Detection rate refers to the ratio between the numbers of correctly detected outliers to the total number of actual outliers. False alarm rate is the ratio between the numbers of normal objects that are misinterpreted as an outlier to the total number of alarms.

Clustering algorithm accuracy is calculated by using the number of the clusters, number of outliers, a time taken by the algorithm, number of instance per cluster, squared error within the cluster, number of iteration required by clustering algorithm, non-clustered instance(if any). Outlier detection accuracy is calculated, to find out the number of outliers which is detected by the clustering algorithms for sample survey data of India (comprises of the Inflation rate, Demonetization, UIDAI, and Pan Card).

We had observed that K-Means algorithm gives the maximum numbers of outliers and DBSCAN algorithm is giving the least number of outliers for the same data set.

5.1 Results of outliers of different algorithms

Table 5.1: Numbers of outlier comparison based on different clustering algorithm on different data samples of India

Algorithms	Sample	No. Of Outliers
K-mean	Demonetization	1178
DBSCAN	Demonetization	127
Single Linkage	Demonetization	617
K-mean	Inflation Rate	456
DBSCAN	Inflation Rate	57
Single Linkage	Inflation Rate	183
K-mean	UIDAI	1876
DBSCAN	UIDAI	627
Single Linkage	UIDAI	1043
K-mean	Pan Card	1351
DBSCAN	Pan Card	493
Single Linkage	Pan Card	784

VI. CONCLUSION

The outlier detection is one of the challenging tasks in data sets. We should take care of exceptional conditions while implementing any policies which can be solved through finding outliers for that particular domain. As discarding these exceptional is

not the solution which may affect the outcome of the policies harshly and shadow the benefits of the policies. In finding the best algorithm to detect the outliers several factors are used. It had observed that K-Means algorithm gives the maximum numbers of outliers and DBSCAN algorithm is giving the least number of outliers for the same data set.

This paper mainly discusses the best algorithm to find the outliers using sample survey data of India. The outliers can be compared to the count of the data points lying outside the quadrant for each of the algorithms for same sample space.

REFERENCES

[1] D.Hawkins. Identification of outliers. Chapman and Hall, Reading, London, 1980

[2] Usama Fayyad, Gregory Piatetsky Shapiro, and Padraic Symyh, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of the ACM, Vol. 39, No. 11, pp. 27-34,1996

[3] Hodge, V. J.; Austin, J. (2004). "A Survey of Outlier Detection Methodologies" (PDF). Artificial Intelligence Review. 22 (2): 85–126. doi:10.1007/s10462-004-4304-y

[4] Chauhan R, Kaur H, Alam MA, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications, (0975 – 8887) Vol.10– No.6, November 2010

[5] Jain A.K., Murty M.N., and Flynn P.J., "Data Clustering: A Review", ACM Computing Surveys, 31 (3). pp. 264323, 1999

[6] Raj Bala, Sunil Sikka, Juhi Singh, " A Comparative Analysis of Clustering Algorithms ", International Journal of Computer Applications (0975 – 8887), Volume 100 – No.15, August 2014

[7] Masciari E., Pizzuti C, and Raimondo G., "Using an out of core technique for Clustering Large Datasets", Proceedings of 12th international workshop of database and expert system Application, Munich, Germany

[8] K. Yamnishi, J. Takeuchi, G. Williams. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. In:Proc of KDD'00,pp.320-325,2000

[9] S.D. Bay, M. Schwabacher. Mining Distance-Based Outliers in Near-Linear Time with Randomization and a Simple Pruning Rule. In: Proc of KDD'03, 2003

[10] M.F. Jiang, S.S. Tseng, C.M. Su. Two-phase Clustering Process for Outliers Detection. Pattern Recognition Letters, 2001, 22(6-7):691-700

[11] C.Aggarwal, P. Yu. Outliers Detection for High Dimensional Data. In: Proc of SIGMOD'01, pp. 3u-46, 2001

[12] L. Wei, W. Qian, A. Zhou, W. Jin, J.X, Yu. HOT: Hypergraph-Based Outlier Test for Categorical Data. In:Proc of PAKDD'03,pp.399-410, 2003

[13] S. Harkins, H. He, G. J. Willams, R. A. Baster. Outlier Detection Using Replicator Neural Networks. In: Proc. of DaWak'02,pp.170-180,2002

[14] G.J. Williams, R.A. Baster, H. He, S. Harkins, L. Gu. A Comparative Study of RNN for Outlier Detection in Data Mining. In: Proc of ICDM'02. pp. 709-712, 2002

[15] H. Liu, K. C. Jezek, and M. E. O'Kelly, "Detecting outliers in irregularly distributed spatial datasets by locally adaptive and robust statistical analysis and gis". International Journal of Geographical Information Science,15(8), 2001. pp.721–741

[16] Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," Proc. Int'l Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169

[17] Jain, A.K., Dubes, R.C., 1988. "Algorithms for Clustering Data". Prentice-Hall Inc

[18] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li," Automated Variable Weighting in kMeans Type Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 27, NO. 5, PP. 657668, 2005

[19] Ester M., Kriegel HP., Sander J., Xu X."A density-based algorithm for discovering clusters in large spatial databases with noise". Second International Conference on Knowledge Discovery and Data Mining (1996)

[20] M. Ankerst, M.M.Breunig, H.-P. Kriegel, J.Sander, "OPTICS: Ordering points to identify the clustering structure", in Proceedings of ACM SIGMOD Conference, 1999 pp. 49-60