Clustering Algorithms for Outlier Detection Performance Analysis

S. Vijayarani,[1] S. Maria Sylviaa ,[2]A.Sakila[3]
[1]Assistant Professor, Dept. of CS, Bharathiar University, Coimbatore, India
[2]M.Phil Research Scholar, Dept. of CS, Bharathiar University, Coimbatore, India
[3]M.Phil Research Scholar, Dept. of CS, Bharathiar University, Coimbatore, India

Abstract—Data mining is the method of extracting the data from large database. Various data mining techniques are clustering, classification, association analysis, regression, summarization, time series analysis and sequence analysis, etc. Clustering is one of the important tasks in mining and is said to be unsupervised classification. Clustering is the techniques which is used to group similar objects or processes. In this work four clustering algorithms namely K-Means, Farthest first, EM, and Hierarchical are analyzedby the performance factors clustering accuracy, number of outliers detected and execution time. This performance analysis is carried out in BUPA (liver disorder) dataset. This work is performed in WEKA data mining tool.

Keywords— Outlier, Clustering, K-means, Farthest first, EM, Hierarchical.

I. INTRODUCTION

Data mining is the method of extracting the data from the vast dataset. It is an essential step of knowledge discovery process by analyzing the massive volumes of data from various perspectives and summarizing it into useful information [3].The most important data mining techniques are classification, clustering, association rule generation and regression [11]. Data mining is used in numerous applications such as medical, stock analysis, fault analysis, forecasting, and science examination. There are numerous task accomplished by data mining they are classification, clustering, association rule mining, prediction, outlier analysis, time series. Clustering plays an important role in data mining process. Clustering is the approach of grouping the data into classes or clusters so that the objects within each cluster have high similarity in comparison with one another[12].

The common approach of clustering techniques is to find cluster centroid and then the data are clustered. Several clustering techniques are partitioning methods [4], hierarchical methods, density based methods, grid based methods [4] and model based methods. Clustering is a challenging field of research in which its potential applications pose their own requirements [4]. Clustering is also called as the data segmentation because clustering method partitions the large data sets into smaller data groups according to their similarities. The main objective of cluster analysis is to increase intra-group similarity and inter-group dissimilarity.Detecting outlier is one of the important tasks in clustering process. A failure to detect outliers or their ineffective handling can have serious ramifications on the strength of the inferences drained from the exercise [4].

Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry and recently for detecting terrorism related activities [5].Outliers are found using the filters which is offered by data mining tools. Liver disorder is also referred to as hepatic disease which damages the liver. Liver is affected mostly due to smoking problem. A number of liver function tests (LFTs) are available to test the proper function of the liver [6]. This test for the presence of enzymes in blood that is normally most abundant in liver tissue, metabolites or products [6]. Interquartile range is the filter used in WEKA tool to find the outlier. Then the data is clustered based on the outlier class and the clustering algorithm such as k-means, Farthest first, hierarchical, EM algorithm are used to find the best algorithm and the performance evaluation is found by changing the number of clusters. The analysis is done using the WEKA (Waikato Environment for Knowledge Analysis) tool and the liver disorder (BUPA) dataset is extracted from UCI repository.

II. REVIEW OF LITERATURE

Manish Verma, et.al [11] have analyzed an overview of six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBscanclustering [11], Density Based Clustering, Optics and EM Algorithm. Banking data related to customer information dataset have been taken for analysis in WEKA tool.Sharmila, et.al [12] have given an overview of four different clusteringalgorithms(K-means, EM, farthest first, hierarchical) and given the advantages and disadvantages of four algorithms.**S.** Revathi, et.al [13] has analyzed the various clustering algorithms. The two different dataset (Letter image, Abalone) have been collected and estimated the time. In this farthest first clustering algorithm takes very few seconds for clustering, whereas the simple K-Means takes the longest time to perform clustering. NamitaBhan, et.al [14] analyzed an in-depth study of k-Means and the Expectation-maximization technique, based on training data and percentage split using the WEKA3.6.9. Fixed broadband internet users dataset have been taken. The EM algorithm shows much better performs than k-means algorithm.Preeti Baser et.al [15] conducted an experimental study of comparative analysis of various clustering techniques and its applications in various domains. Authors also focused on the classification of clustering techniques.

III. METHODOLOGY

Information retrieval is the method used to collect the data from the large database. Using Interquartile range the outlier is found among the whole instances. By using k-means, farthest first, EM, Hierarchical clustering algorithm the clustered and un-clustered instances are found based on number of clusters. The best algorithm is also found for retrieving information from the bupa dataset. The process flow of comparative analysis is illustrated in Figure 1.

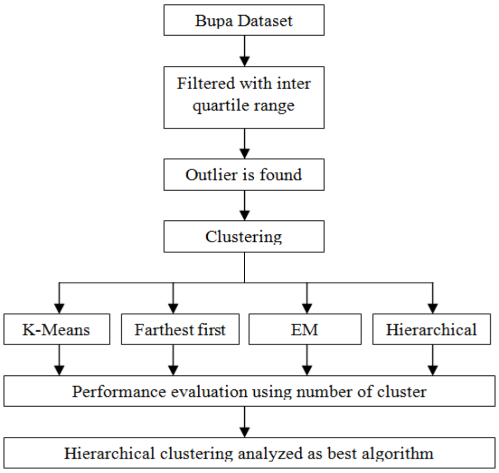

Bupa Dataset
Filtered with inter quartile range
Outlier is found
Clustering
K-Means
Farthest first
EM
Hierarchical
Performance evaluation using number of cluster
Hierarchical clustering analyzed as best algorithm

Figure 1: Proposed Methodology

A. Dataset

The BUPA (liver disorder) dataset is collected from UCI (UC Irvine) repository. The dataset contains 345 instances and 7 attributes. The machine learning data mining tool WEKA is used to find the outlier in the dataset and to find the performance of algorithm based on the number of clusters. The outlier is found using the interquartile range filter and the attribute such as outlier and extreme values are added to the original attributes. Here, after the usage of filter 9 attributes are defined.

B. Overview of Liver Disorder

Liver is the largest organ in human body.The main functions of liver are to accumulate and metabolize nutrient, fabricate protein, detoxify blood. Alcoholic liver disease is known as etiology but a complex and incompletely known pathogenesis and it is a disease with particular morbidity and mortality. Alcoholic liver disease is the result of excessive alcohol consumption and is seen in all social groups including those who consume regularly but who are not addicted to alcohol [9].There are different liver disorder they are hepatitis, liver tumor.

C. Outlier

The abnormal data in the dataset defined to be the outlier. These patterns are referred as noise, outliers, faults, novelty, errors, surprise, exceptions, peculiarities [7], observations defects, contaminants and aberrations in distinct application domain [7]. The various domains such as master card, medical, intervention detection use the outlier technique for detection [7].In WEKA, interquartile range is the filter used to detect the outlier. The malicious intentions are performed inside the system by the intruders that are detected as outliers and by using the detection techniques the faults can be detected by monitoring every activity of the data [8]. The interquartile range (IQR) is the collection of the values of a variable over the focus part of a distribution and it is the choice from the 25th to the 75th percentile of a variable.It is calculated using

$$\boxed{IQR = P75 - P25}$$

D. Clustering

Clustering technique is used to divide the data into number of groups. The data with similar characteristic will form a cluster. Clustering is one of the important tasks in mining and is said to be unsupervised classification. Clustering is challenging fields of research in which its potential applications pose their own requirements [4].Clusters are composed based on the centre point known as centroid. There are various algorithms used to cluster the data.

E. K-Means

K-Means is an iterative clustering algorithm in which the items are moved among the set of clusters until the desired set is reached [13]. In almost all cases, the simple K-Means clustering algorithm takes more time to form clusters [13].It will partition the n observational data into k clusters, in this each data is present in one of the clusters.

Table 1

```
Algorithm K-Means (k, D)

1. Chooses k data points as the initial centroids
(cluster Centers)
2. Repeat {
3. for each data point x ∈D do {
4. Compute the distance from x to each centered;
5. Assign x to the closest centered // a centered
represents a cluster
6. end for }
7. Re-compute the centered using the current
cluster
memberships
8. Using till the stopping criterion is met }
```

F. Farthest First

Farthest first is an alternative of K Means algorithm. Farthest first places the cluster centroids from that the other clusters are placed. Thesecentroids must lie within the data area. The points that are farther are clustered together and this feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment [13].

Table 2

```
ALGORITHM Farthest-k-Object
```

```
1. Initialize InitCent=Φ;
2. Randomly choose any of the objects from $D_{trainset}$ as first point $C_{init1}$
3. InitCent=InitCent {$C_{init1}$}
4. $C_{init2}$=Farthest object O Є $D_{trainset}$ from $C_{init1}$
5. InitCent=InitCent U {$C_{init2}$}
6. If K=2, return InitCent; exit; Else
7. i=3;
8. Repeat
9. $C_{init1}$=Object O Є $D_{trainset}$, such that sum of its distance from all the points in InitCent is maximum;
10. InitCent=InitCent U {$C_{init1}$}
11. i=i+1;
12. Until i=k
13. End
```

G. EM

EM is said to be expectation maximization algorithm it is a continuous method of finding the maximal likelihood. Maximum likelihood is the scheme of finding the parameters. The EM iteration alternates between performing an expectation (E) step [16], which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters and maximization (M) step which computes parameters maximizing the expected log-likelihood found on the *E* step [16].

Algorithm EM
1. Initialize: Set i=1 and choose an initial $\theta_{1.}$
2. While not converged do:

a. Expectation (E) step: Compute

$$\varphi = \sum_{i=1}^{N} E_{\theta i}\{X^2{}_t|Y\},$$

$$\psi = \sum_{i=1}^{N} E_{\theta i}\{X_tX_t+1|\},$$

b.Maximization (M) step: Find the next iterate according

$$\theta_{i+1} = \frac{\psi}{\varphi}$$

c.If $|L_{\theta i}(Y)-L_{\theta i-1}(Y)| \geq 10^{-6}$, update i:=i+1 and terminate.

H. Hierarchical

Hierarchical clustering algorithms (HAC) are either top down or bottom up. Bottom- up algorithms treats each document as a singleton cluster at the outset and then successively merges pairs of clusters until all clusters have been merged into a single cluster that contains all documents [10]. Bottom up hierarchical clustering is known as HAC. Top-down clustering requires a method for splitting a cluster and HAC proceeds by splitting clusters recursively until individual documents are reached [10].

Table 3

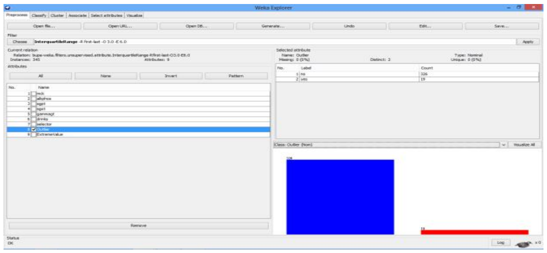
```
Algorithm Hierarchical (n, C)
1. Start with n clusters containing one object
2. Find the most similar pair of clusters Ci,eCj from the proximity matrix and merge them into a single cluster
3. Update the proximity matrix (reduce its order by one, by replacing the individual clusters with the merged cluster)
4. Repeat steps (2) e (3) until a single cluster is obtained (i.e. N-1 times)
```

IV. Experimental Results

A. Detecting Outlier

In this paper the interquartile range filter is used to find the outlier and then four clustering algorithms have been compared to find the better performance based on the number of clusters in the bupa dataset which contains 345 instances and 7 attributes after applying the filter the number of attribute increases to 9 attributes.Figure2shows the total number of outliers detected.

Figure 2:Total number of Outliers Detected



Filter
InterquartileRange -R first-last -O 3.0 -E 6.0

Current relation
Relation: bupa-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-O3.0-E6.0
Instances: 345 Attributes: 9

Attributes

No.	Name
1	☐ mcv
2	☐ alkphos
3	☐ sgpt
4	☐ sgot
5	☐ gammagt
6	☐ drinks
7	☐ selector
8	☑ Outlier
9	☐ ExtremeValue

Selected attribute
Name: Outlier Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	no	326
2	yes	19

Class: Outlier (Nom)

Status
OK

B. Cluster Accuracy

Cluster accuracy can be found by calculating the clustered and un-clustered instances. The clustered instances can be found by changing the number of clusters based on the outlier class they are shown in Figure 3.Un-clustered instances are shown in Figure 4 which defines the accuracy of clustering. In Table 1 the clustered, un-clustered instances and time taken for clustering the data have been shown which are based on the number of clusters. The number of clusters (2, 3, 4, and 5) has been changed for finding the accuracy for every clustering algorithm.Figure 5 shows the time taken for building the model.

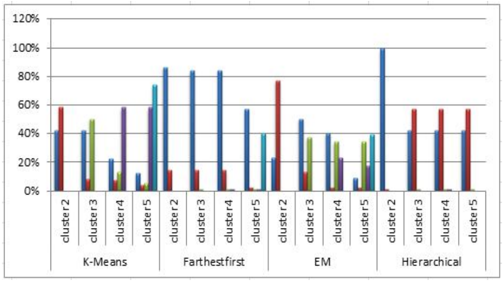


Figure 3: Clustered instance

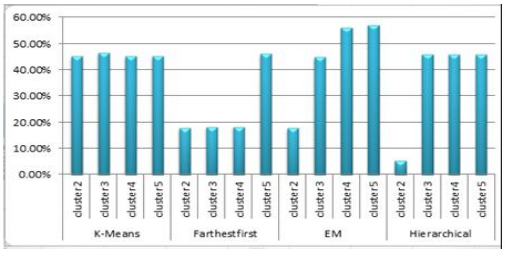


Figure 4: Un-Clustered instance

Table 4: Clustered, Un-clustered and Time Taken

Algorithms	K-Means				Farthest first				EM				Hierarchical			
	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
Clustered Instance	145	145	75	40	295	290	290	195	78	173	139	31	340	145	145	145
	200	26	24	15	50	50	50	8	267	45	7	6	5	195	195	195
		74	46	16		5	3	3		127	118	116		5	3	3
			200	200			2	2			81	58			2	1
				74				137				134				1
Un-Clustered Instance	156	160	156	156	61	63	63	159	61	155	194	197	18	158	158	158
No. of Outliers	4	4	0	0	15	12	12	12	18	1	1	1	16	4	4	4
	15	13	4	4	4	4	4	3	1	18	5	4	3	12	12	12
		2	0	0		3	2	2		0	0	0		3	2	2
			15	15			1	1			13	14			1	0
				0				1				0				1
Time taken	.03	.02	0	0	0	0	0	0	.06	.11	.14	.19	.02	.02	.19	.17

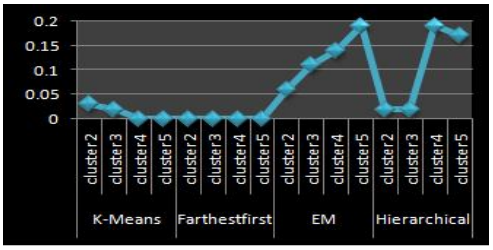


Figure 5:Time taken

V. CONCLUSION

Clustering techniques are the unsupervised methods which are used to organize the data into groups based on their similarities. Most clustering algorithms do not depend upon assumptions that match to the traditional statistical methods such as the distribution of the statistical data [1].Running the clustering algorithm using any software produces almost the same result even when changing any of the factors because most of the clustering software uses the same procedure in implementing any algorithm [11].In this paper the outlier is found by the help of the filter and then the number of clusters is divided to find the outlier accuracy, time taken to build and the performance is evaluated by comparing four algorithms.Hierarchical clustering algorithm is more sensitive for noisy data [11]. From this analysis it is found that the hierarchical clustering algorithm performance is efficient and accurate.

REFERENCES

[1] A.J.Patil, C.S.Patil, R.R.Karhe, M.A.Aher- "Comparative Study of Different Clustering Algorithms" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 3, Issue 7, July 2014.

[2] KhaledAlsabti, Vineet Singh, Sanjay Ranka-"An Efficient K-Means Clustering Algorithm".

[3] AhamedLebbe Sayeth Saabith, ElankovanSundararajan, Azuraliza Abu Bakar –"Comparative study on differentclassification techniques for breast cancer dataset" International Journal of Computer Science and Mobile Computing, Vol.3 Issue.10, October- 2014.

[4] S.VijayaraniS.Nithya –"An Efficient Clustering Algorithm for Outlier Detection" -International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011 22

[5] Ajay Challagalla,S.S.ShivajiDhiraj ,D.V.L.N Somayajulu,TomsShajiMathew,SauravTiwari,SyedSharique Ahmad " Privacy Preserving Outlier Detection Using Hierarchical Clustering Methods,2010 34th Annual IEEE Computer Software and Applications Conference Workshops.

[6] http://en.wikipedia.org/wiki/Liver_disease

[7] Varunchandola, arindambanerjee and vipinkumar-" Outlier Detection: A Survey"-, TR 07017 August 15,2007

[8] Victoria J. Hodge and Jim Austin-"A Survey of Outlier Detection Methodologies"-, Springer 2004

[9] Liver Disease in Canada, "A Crisis in the making" an assessment of liver disease in canada march 2013

[10] "Hierarchical clustering" Online edition Cambridge up draft! © April 1, 2009 Cambridge University Press. Feedback welcome. 377

[11] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta, ISSN: 2248-9622 "A Comparative Study of Various Clustering Algorithms in Data Mining" International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012.

[12] Sharmila, R.C Mishra, ISSN: 2231-5381 "Performance Evaluation of Clustering Algorithms" International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013.

[13] S. RevathiDr.T.Nalini "Performance Comparison of Various Clustering Algorithm" International Journal of Advanced Research in Computer Science and Software Engineering Research Paper,Volume 3, Issue 2, February 2013.

[14] NamitaBhan, (Dr.) DeeptiMehrotra "Comparative Study OfEm And K-Means Clustering Techniques In Weka Inter-Face" International Journal of Advanced Technology & Engineering Research (IJATER), Volume 3, Issue 4, July 2013.

[15] Preeti Baser,Dr. Jatinderkumar R. Sain, "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets" International Journal of Computer Science & Communication Networks,Vol 3(4),271-275 271

[16] http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm