ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

Improving Cluster Formulation to Reduce Outliers in Data Mining

Nancy Lekhi[1], Manish Mahajan[2]

M.Tech Student, CEC Landran, Punjab, India.[1]

Associate Professor, CEC Landran, Punjab, India[2]

ABSTRACT: Existing studies in data mining focus on Outlier detection on data with single clustering algorithm mostly. There are lots of methods available in data mining to detect the outlier by making the clusters of data and then detect the outlier from them .Where outlier is the data item whose value falls outside the bounds in the sample data may indicate anomalous data. Outlier can be reduced if we improve the clustering .In this paper we proposed a hybrid algorithm that work not only on numeric data but also on text data. Our focus is to improve the cluster making so that the number of outliers can be reduce for that we can combine the clustering and classification techniques of data mining i.e. weighted k-mean and neural networks.
.

KEYWORDS: energy Data Mining; Outlier; Clustering; k-means; weighted k-mean; Neural Networks

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amount of data .Data mining refers to fetching hidden and necessary information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality [6]. Data mining consist of lots of techniques i.e. cluster detection, Decision Tress, Memory based reasoning, link analysis, Neural networks, Genetic algorithms and lot more.

A data items whose values are different from rest of data or whose values falls outside the described range are called outlier. Or you can say that database may contain data that do not comply with the general behaviour or model of rest of the data. An outlier is an observation that deviates from other observations as to arouse suspicion that it was generated by a different mechanism [1].

Clustering is the earliest data mining technique. This technique designed as undirected knowledge discovery or unsupervised learning. There are lot of clustering techniques which are used to generate the clusters from data.

II. RELATED WORK

In [2] author presented a simple and efficient implementation of k-means clustering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. a popular k-means algorithm organizes or groups data by firstly assigning all data points to the nearest clusters, then determining the cluster centre. The algorithm repeats these two steps until it has divide the data into groups that's why the k-means algorithm speed in making clusters is very slow. in this [3] author shows the increase in speed and accuracy of k means by add a variation to the k-means to do better with a large data set called weighted k-means and having much difference in cluster density to enhance the clustering scalability. To speed up the clustering process, the author develops the reservoir-biased sampling as an efficient data reduction technique. But this method does not work on real large databases and there is not any method for outliers. In this paper [4] the author discussed major problem of using k-mean type algorithms in data mining is selection of variables (attributes). K-mean algorithm cannot select variables automatically because they treat all variables equally in the clustering process that result in poor clustering. New k-mean type clustering algorithm called weighted-k-mean that can automatically calculate variable weights. But this algorithm's are week or poor to find the outlier. In [5] author has developed an new algorithm for outlier detection using genetic algorithm.Genetic algorithm is better in computing the number of outliers in a particular time period. But this method does not work on

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

dataset of various types and required to improve the processing speed and performance of the algorithm. The author in [6] enhanced the work done by uses hybrid approach for outlier detection the principle of outliers finding depend on the threshold. Threshold is set by user. But This approach is only deals with numerical data not with text data or on mixture data and also performance of this approach is low. In this paper [7] author presented an algorithm that provides outlier detection and data clustering simultaneously. In this the author used two technique one is for clustering i.e. genetic k means and other for outlier detection i.e. outlier removal clustering. But this can work for large scale data of same type not for mixed type.

III. CLUSTERING TECHNIQUES

Clustering algorithm searches for groups or clusters of data elements that are similar to one another..Principle of clustering is to Maximizing intra-class similarity & minimizing inter-class similarity as shown in figure 1.

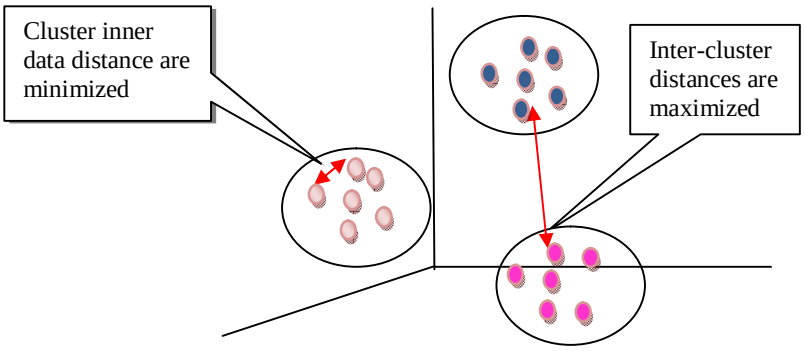

Fig 1: Clustering

There are lot of clustering techniques which are used to generate the clusters from data. Let we discuss some of them:-

K-mean: - The simplest and most commonly used algorithm to make a cluster is the K-means algorithm. This algorithm divides the data into K clusters i.e. C1 to CK represented by their centers [2]. The center of clusters is calculated as the mean of all the instances belonging to that cluster. K-mean algorithm will do three steps:-

Input: n data points and the number of cluster (*K*)
Output: *K* clusters

i. Initialize the *K* cluster centers
ii. while termination condition is not satisfied do
Determine the distance of each object to the centroids
Group the object based on minimum distance (find the closest centroid)
iii. end while

Disadvantage: - the main disadvantage of k-mean algorithm is that it only works on numeric data not on text or any mixed type of data.

Weighted k-mean:- The proposed extension to the k-means algorithm is called weighted k-means[3].This algorithm is very useful method and it overcomes the disadvantages of k-mean algorithm i.e. it work on numeric as well as text and mixed type of data (date and time).

Input: n data and the number of cluster (*K*)
Output: *K* clusters

i. Initialize the k cluster center
ii. For loop until all data is processed
a. Randomly generate the weights for n number of data
b. And count distance from randomly generated centers
c. End for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

iii. Now divide the data having weights into k clusters.

IV. CLASSIFICATION TECHNIQUES

The objective of classification is to analyze the input data and to develop an accurate Description or model for each class using the features present in the data. There are some Classification Techniques that used in data mining:-

Genetic Algorithms (GA):-Genetic Algorithm is used to generate useful solutions to optimization and search problems . GA is an algorithm which makes it easy to search a large search space [7]. But GA has many drawbacks like GA take so much time and they cannot always find the exact solution but they always find best solution. And it work well with numeric data.GA works on large amount of data effectively but not on small amount of data.

Neural Network (NN):- NNs classify/recognize existing patterns based on training that you provide. , NN is a mathematical model or computational model based on biological neural networks [8]. NN are models of intelligence that consist of large numbers of simple processing units collectively are able to perform very complex pattern matching tasks. NN is fast technique as compare to GA and its results are always better and also it has high accuracy. Even it work on small amount of data as well as large amount of data so efficiently.

V. PROPOSED ALGORITHM

As we discuss earlier that we would able to reduce the outlier by performing better clustering. This method of clustering could formed through combining two techniques together so we could able to create the system that make an output of one technique that could introduce as input of the other technique. So that system could perform better and exclusively enhance in performance

A. *System Architecture*

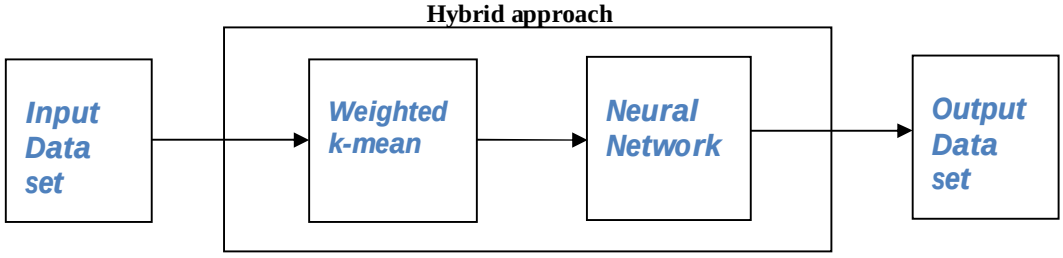

Fig 2: System Architecture

- ***Input Data Set*:** Collecting dataset for input.
- ***Weighted k-mean:*** Weighted k-mean is a method of clustering. It provides the weights to each data element and then divides them into k number of clusters. Results of weighted k-mean will save by user.
- ***Neural Networks:*** Neural network contain number of neurons that work on data and generate the better results .The output of weighted k-mean that is stored by user is then given to the neural as input after that neural generate the output dataset.
- ***Output Data Set:*** Contains the accurate clusters of input dataset.

B. *Proposed Weighted k-mean and Neural Network Algorithm*

Aim of the proposed algorithm is to improve the data clustering by combining two techniques together so that number of outlier reduces. The proposed algorithm is consists of two main steps.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

Step 1: Calculating the distance of elements:

The location of each element is calculated randomly by eq. (1) i.e. round off and random function.

$$x_2=\texttt{round(n*rand)}$$
$$y_2=\texttt{round(n*rand)}$$

eq. (1)

Where each time the random number is generate and then multiply by n where n can be any number. And center of total elements location is calculated by using eq. (2)

$$x_1=\texttt{mean(x)}$$
$$y_2=\texttt{mean(y)}$$

eq. (2)

where x and y are the total number of locations now calculate the center of each x and y. After that calculate the distance of each element by using eq. (3)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

eq. (3)

Step 2: Assign the weights to each element and calculate the average of weights

Assign the distance as their weights to each element and then calculate the average of that weights by using eq.(4)

avg_wt=mean (d)

eq. (4)

where d is the distance that we calculated before .

VI. PSEUDO CODE

Step 1: Browse the data file that contains text, numbers and date/time.
Step 2: Calculate x_2 using eq. (1). And y_2 using eq. (2).
Step 3: Calculate the distance by putting the values in eq. (3) that is obtained from eq.(1) and eq.(2) .
Step 4: Assign the distance as a weights to each element.
Step 5: Calculate the average of weights by using eq. (4).
Step 6: repeat the below steps until the all elements are processed
Check the below condition for making the clusters
If (d<=avg_wt)
assign first_clust=d
else
assign second_clust=d
end
end loop
Step 7: Save the output of Step 6 that could introduce as input of the Neural Network technique.
Step 8: End.

VII. SIMULATION RESULTS

The proposed algorithm is implemented with MATLAB. In fig 3. The text data is shown by red points which are divided into two clusters using weighted k-mean Clustering Technique. Then the results of clustering is pass as input to neural network for classifications after providing the training to neural the testing results are obtained fig.4 shows the Neighbor Weight Distances and fig.5. Shows the weights from two input where the dark area indicates the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

high weights and low area indicate the low weights. Where fig. 6 shows how many data points are associated with each neuron here the data are evenly distributed across the neurons. Fig. 7. Shows the locations of the data points and the weight vectors.

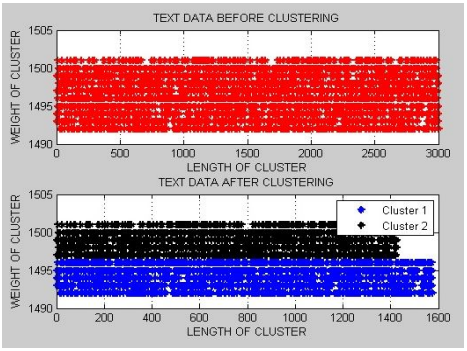

Fig. 3. Clustering perform on text data

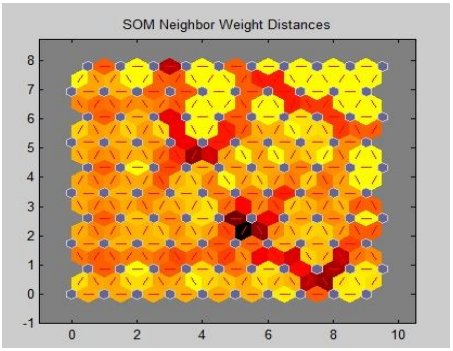

Fig 4. Neural generate the Neighbour Weight Distance

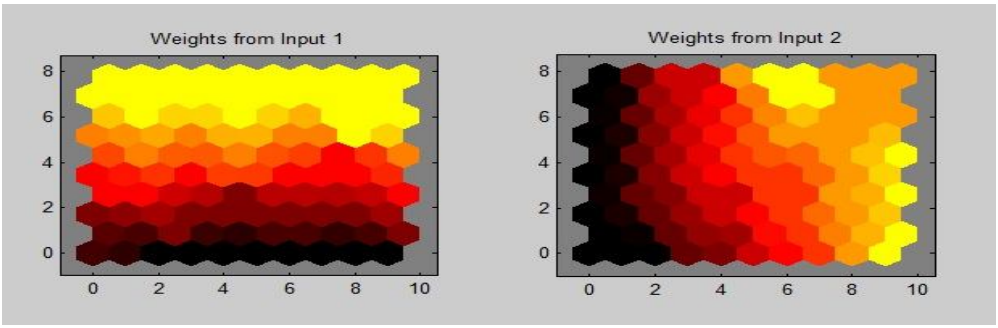

Fig. 5. Neural shows the weights of text Input

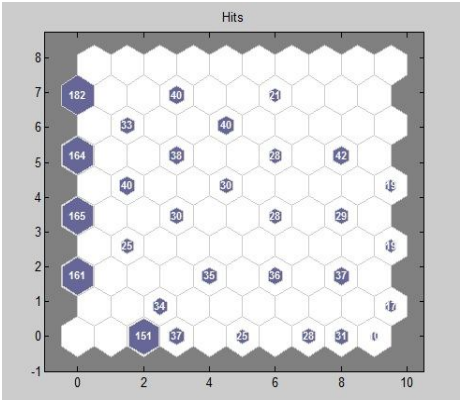

Fig. 6. SOM Number of hits per neurons

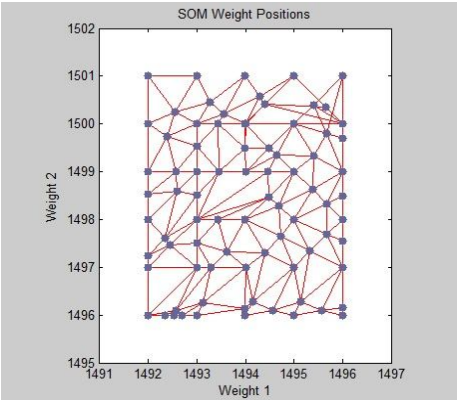

Fig 7. Plot SOM Weight Positions

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

VIII. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better than that of genetic k-mean. It also increases the classification accuracy using weighted k mean clustering algorithm. This proposed method deal with text dataset that has not been implemented before using genetic k-mean on the text dataset but rather performed on numeric dataset. We have used text and numeric data only. The future aspects of this research work might involve performing the clustering process on compound dataset to analyze the performance.

REFERENCES

1. D.Hawkins: "Identification of outliers". Chapman and Hall, London. 1980.
2. Tapas Kanungo, Nathan S. Netanyahu, Angela Y "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002
3. Kittisak Kerdprasop, Nittaya Kerdprasop, and Pairote Sattayatham "Weighted K-Means for Density-Biased Clustering"A Min Tjoa and J. Trujillo (Eds.): DaWaK 2005, LNCS 3589, pp. 488-497, 2005.Springer-Verlag Berlin Heidelberg 2005
4. Anand M. Baswade,Kalpana D. Joshi,Prakash S. Nalwade "A Comparative Study Of K-Means And Weighted K-Means For Clustering" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181
5. P. Vishnu Raja,Dr. V. Murali Bhaskaran "An Effective Genetic Algorithm for Outlier Detection" International Journal of Computer Applications (0975 – 8887) Volume 38– No.6, January 2012
6. Ms. S. D. Pachgade, Ms. S. S. Dhande "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach"International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, June 2012
7. M. H. Marghny , Ahmed I. Taloba , " Outlier Detection using Improved Genetic K-means " IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. 38, NO. 11, July 2013
8. DR. YASHPAL SINGH, ALOK SINGH CHAUHAN "Neural Networks In Data Mining" Journal of Theoretical and Applied Information Technology © 2005 - 2009 JATIT

BIOGRAPHY

Nancy Lekhi is a Research Assistant in the Information Technology Department, Chandigarh Engineering College, Punjab Technical University. She received her B.Tech in Information Technology from Baba Banda Singh Bahadur College of Engineering in 2012, India. Her research interest is Data Mining.