Entropy Reduction By Using K-Means Clustering And Neural Network

Ms. Pankti Dubal[1], Ms. Disha Sanghani[2]
Department of I.T
[1]Research Scholar,GTU, Gujarat
[2]Assistant Professor,SSEC, Bhavnagar, Gujarat

Abstract*-Vast amount of data is being gathered by numerous medical organization. Researchers are working hard to uncover hidden patterns from raw data available. This paper focuses on the tools and various data mining techniques to study the medical data. Classification and clustering are two major task done in data mining. In clustering partition of the data set in done in homogeneous groups based on some specific features. There are many clustering techniques available today. Based on the dimension of the data, certain techniques work efficiently for low dimensional and fails to handle high dimensional data. There are many clustering algorithms such as K- means and hierarchical, from which hierarchical clustering is of great importance in real world data because data size is increasing exponentially. In this paper, we have proposed an algorithm combining various clustering algorithm and making a hybrid clustering. We have tried to improve efficiency by reducing computational time even when data dimensionality is huge. Later the classifier is used to label the unlabeled data obtained from hybrid clustering. Labelling of the data can be done by using neural network.*

Keywords-K-means, Neural Network, entropy, clustering, Classification.

I. INTRODUCTION

Researchers are working very hard nowadays to achieve fast and efficient algorithm that can abstract medical data. Cancer has become a leading cause of death in India. There are various sites of cancer in our body such as blood, oral, stomach and bone cancer and many more. Cancer is curable if predicted at early stage. As huge amount of data is available from medical organization we need efficient technique that can recognize the hidden pattern from huge data available. Proposed work will focus on studying medical pattern detection and predict cancer as soon as possible. Clustering is an unsupervised technique of classification of pattern into cluster.

Huge amount of data is available from various information system. So, every day we are left with huge amount of raw data. Clustering is applied on these data for medical decision-making. Later section discusses various clustering techniques.

The Paper is organized as follows. Section II provides a brief discussion on the previous works related to the topic. Section III explains the distance measure techniques. Section IV discuss the problem statement. The proposed methodology is presented in Section V, while Section VI presents the workflow of the proposed system. And Section VII presents the experimental results. A brief summary along with future research directions is given Section VII.

1. K- means:

Clustering is a data mining technique used to cluster the observation into clusters without prior knowledge about the data. The main reason behind popularity of k-means is that its time complexity is linear. The complexity for M iteration on data set having N instance with A attributes in each instance is about O(K * M * N * A).[9]

1) Pick a number (K) of cluster centers
2) Assign every data point (e.g., gene) to its nearest cluster center
3) Move each cluster center to the mean of its assigned data points.
4) Repeat 2-3 until convergence

The main disadvantage of k-means technique is that the k value for making clusters are decided by the users. The technique is not efficient to select the appropriate value of k depending upon the data sets. This method works well with nominal attributes.

2. Hierarchical Clustering:

The output of hierarchical algorithm is in form of a dendogram. Dendogram is tree representation of the data. The hierarchies is represented by different levels in the tree.

Data at different hierarchy are combined based on similar pattern to form a bigger one. Hierarchical clustering involves cluster arranged in form of a tree in data mining.

The hierarchical clustering can be done in two ways as mentioned below:

a. Agglomerative hierarchal clustering:

This is also called “bottom up” approach. In this method distance between different pairs of cluster is calculated. The cluster with minimum distance is combined and similarly this process continues until a single cluster is obtained. [5]

1. Calculating the proximity matrix for the initial clusters.
2. Searching for the minimal distance in the matrix.
3. Combining the two clusters with the minimal distance.
4. Updating the proximity matrix by calculating the distances between the new cluster with the other clusters
5. Repeating the previous three steps if more than one cluster remains.

3. Divisive hierarchical clustering:

This is also called “top-bottom” approach. This method begins with a single cluster and gradually fragment into smaller clusters until termination condition is reached.

Hierarchal clustering can further be divided into two categories:

1. Single-link clustering: This method is also called nearest neighbor method. It calculates distance between all the elements of two cluster. Then the minimum distance between elements of one cluster to element of another cluster is the distance between two considered clusters.

2. Complete-link clustering: This method is also called furthest neighbor method. This method calculates the distance between two cluster equal to

3. Average-link clustering: This method is called known as minimum variance method. This method calculates the distance between two cluster equal to the average distance between an element of one cluster to any element of other cluster.

II. DISTANCE MEASURE

An efficient distance measuring function plays an important role in data mining techniques. Almost all the clustering algorithm depends on distance function to obtain output. The distance function is used to measure distance between pairs of elements in data set. Many information retrieval techniques use distance function to find the data points that are similar to considered query.K nearest neighbor classifier depends on distance measure to identify the nearest neighbor for data classifications. So, learning distance functions is very important in both data mining and machine learning.

The two common distance measure techniques used in data mining are
1. Euclidean Distance
2. Manhattan Distance

1. Euclidean Distance:

Consider two points in two-dimensional space (a1,a2) and (b1,b2) , the distance between these two points can be calculated using Euclidean formula as shown below:

$$d(a,b) = \sqrt{(a1 - b1)^2 + (a2 - b2)^2}$$

The same concept can be extended to multi-dimensional space. The distance between points can be given by the formula

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)}$$

2. Manhattan Distance:

The Manhattan distance function calculates the distance between two data points if we had to follow grid like path. Manhattan distance between two data points is the sum of the difference of their corresponding coordinates.

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Where n is the number of attributes, and x_i and y_i are the values of the *i*th variable, of points *X* and *Y* respectively.

The significance of Manhattan distance over Euclidean distance is that, Manhattan distance considers the absolute value of the distance to be calculated whereas Euclidean distance doesn’t use absolute values. Using absolute values doesn’t make much changes in the final output even if there are unusual attributes present during distance calculation

Unusual attributes present during distance calculation.

III. PROBLEM STATEMENT

Outliner detection is an important task to be done while studying the data. Outline detection mainly concentrate on removing the dissimilar data from the whole data set. It is used in many fields like fraud detection, network intrusion and diagnosis of diseases. In data analysis applications, outliers are often considered as error or noise and are removed once detected.

In order to detect the irrelevant data from the data set we can use any data mining algorithm. Let us use clustering method, which divides the data set into clusters. Using these clusters we can construct a minimum spanning tree (MST).From this tree, the subtree with minimum number of nodes was considered as outliners and was removed.

IV. METHOLODOLOGY

Partitioning Around Medoids or the K-medoids algorithm is a partitioned clustering algorithm which is slightly modified from the K-means algorithm. They both attempt to minimize the squared-error but the K-medoids algorithm is more robust to noise than K-means algorithm. In K-means algorithm, they choose means as the centroids but in the K-medoids, data points are chosen to be the medoids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal.

The difference between k-means and k-medoids is analogous to the difference between mean and median: where mean indicates the average value of all data items collected, while median indicates the value around that which all data items are evenly distributed around it. The basic idea of this algorithm is to first compute the K representative objects which are called as medoids. After finding the set of medoids, each object of the data set is assigned to the nearest medoid. That is, object i is put into cluster vi, Shift-out membership: an object pi may need to be when medoid mvi is nearer than any other medoid mw.

The algorithm proceeds in two steps:

- BUILD-step: This step sequentially selects k "centrally located" objects, to be used as initial medoids
- SWAP-step: If the objective function can be reduced by interchanging (swapping) a selected object with an unselected object, then the swap is carried out. This is continued till the objective function can no longer be decreased.

The algorithm is as follows:

1. Initially select k random points as the medoids from the given n data points of the data set.

2. Associate each data point to the closest medoid by using any of the most common distance metrics.
3. For each pair of non-selected object h and selected object i, calculate the total swapping cost TCih.

If TCih< 0, i is replaced by h

1. Repeat the steps 2-3 until there is no change of the medoids.

There are four situations to be considered in this process:

I. Initially shifted from currently considered cluster of oj to another cluster;

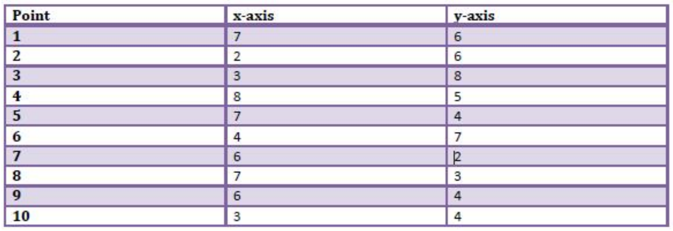
II. Update the current medoid: a new medoid oc is found to replace the current medoid oj ;

III. No change: objects in the current cluster result have the same or even smaller square error criterion(SEC) measure for all the possible redistributions considered;

IV. Shift-in membership: an outside object pi is assigned to the current cluster with the new (replaced) medoid oc.

Example:

For a given k=2, cluster the following data set using PAM.



Point	x-axis	y-axis
1	7	6
2	2	6
3	3	8
4	8	5
5	7	4
6	4	7
7	6	2
8	7	3
9	6	4
10	3	4

Let us choose that (3, 4) and (7, 4) are the medoids. Suppose considering the Manhattan distance metricas the distance measure,

So, now if we calculate the distance from each point:

For (7, 6), Calculating the distance from the medoids chosen, this point is nearest to (7, 4)
For (2, 6) , Calculating the distance from the medoids chosen, this point is nearest to (3, 4)

For (3, 8) , Calculating the distance from the medoids chosen, this point is at same distance from both the points. So choosing that it is nearest to (3, 4)

For (8, 5) , Calculating the distance from the medoids chosen, this point is nearest to (7, 4)
For (4, 7) , Calculating the distance from the medoids chosen, this point is nearest to (3, 4)
For (6, 2) , Calculating the distance from the medoids chosen, this point is nearest to (7, 4)
For (7, 3) , Calculating the distance from the medoids chosen, this point is nearest to (7, 4)
For (6, 4) , Calculating the distance from the medoids chosen, this point is nearest to (7, 4)

So, now after the clustering, the clusters formed are: {(3,4), (2,6), (3,8), (4,7)} and{(7,4), (6,2), (6,4), (7,3), (8,5), (7,6)}.

Now calculating the cost which is nothing but the sum of distance of each non-selected point from theselected point which is medoid of the cluster it belongs to.

Total Cost = cost ((3, 4), (2, 6)) + cost ((3, 4), (3, 8)) + cost((3, 4), (4, 7)) + cost((7, 4), (6, 2))+ cost((7, 4), (6, 4))+ cost((7, 4), (7, 3))+ cost((7, 4), (8, 5))+ cost((7, 4), (7, 6))
= 3 + 4 + 4 + 3 + 1 + 1 + 2 + 2
= 20.

So, now let us choose some other point to be a medoid instead of (7, 4). Let us randomly choose (7, 3).

Not the new medoid set is: (3, 4) and (7, 3). Now repeating the same task as earlier:

So, now if we calculate the distance from each point:

For (7, 6), Calculating the distance from the medoids chosen, this point is nearest to (7, 3)
For (2, 6) , Calculating the distance from the medoids chosen, this point is nearest to (3, 4)
For (3, 8) , Calculating the distance from the medoids chosen, this point is nearest to (3, 4)
For (8, 5) , Calculating the distance from the medoids chosen, this point is nearest to (7, 3)
For (4, 7) , Calculating the distance from the medoids chosen, this point is nearest to (3, 4)
For (6, 2) , Calculating the distance from the medoids chosen, this point is nearest to (7, 3)
For (7, 4) , Calculating the distance from the medoids chosen, this point is nearest to (7, 3)
For (6, 4) , Calculating the distance from the medoids chosen, this point is nearest to (7, 3)

Calculating the total cost = cost((3, 4), (2, 6)) + cost((3, 4), (3, 8)) + cost((3, 4), (4, 7)) + cost((7, 3), (7, 6)) + cost((7, 3), (8, 5)) + cost((7, 3), (6, 2)) + cost((7, 3), (7, 4)) + cost((7, 3), (6, 4))
= 3 + 4 + 4 + 3 + 3 + 2 + 1 + 2
= 22.

The total cost when (7, 3) is the medoid > the total cost when (7, 4) was the medoid earlier. Hence, (7, 4) should be chosen instead of (7, 3) as the medoid. Since there is no change in the medoid set, the algorithm ends here. Hence the clusters obtainedfinally are: {(3,4), (2,6), (3,8), (4,7)} and{(7,4), (6,2), (6,4), (7,3), (8,5), (7,6)}.

V. NEURAL NETWORK AS CLASSIFIER

Neural network [1] is an interconnected group of nodes, a kind of vast network of neurons in a brain. It is used for pattern recognition. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another. Neural network is of two types-

1. Artificial method.
2. Feed forward method

Advantages of using neural-

1. High tolerance of noisy data
2. Can classify the data on which it has not been trained.
3. Classifier that can reduce entropy effectively.
4. Takes the input of clustering algorithm.

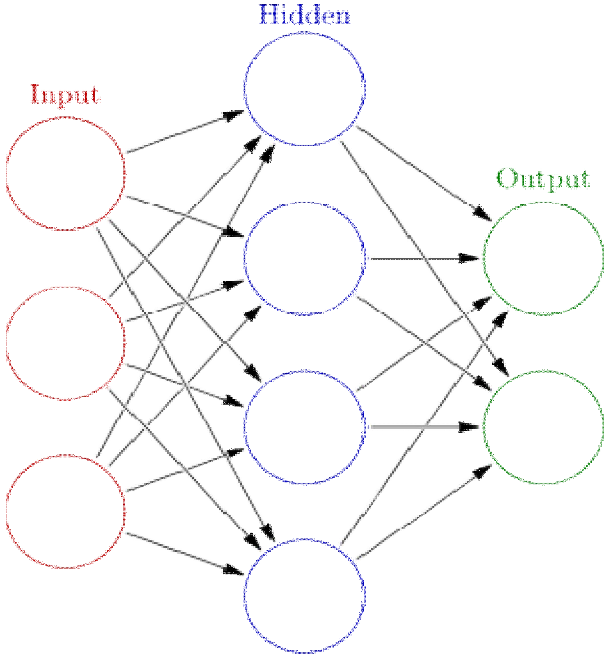


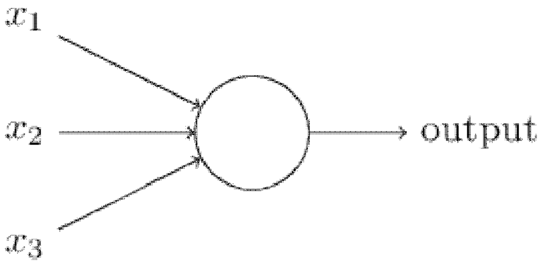
Figure: 1 Neural Network.

Characteristics of Artificial Neural Networks

A large number of very simple processing neuron-like processing elements

A large number of weighted connections between the elements

A perceptron takes several binary inputs, x_1,x_(2 ,)…, and produces a single binary output:



In the example shown the perceptron has three inputs, x_1, x_2, x_3. In general it could have more or fewer inputs. Rosenblatt proposed a simple rule to compute the output.

He introduced *weights* $w_1, w_{2,........}$, real numbers expressing the importance of the respective inputs to the output. The neuron's output, 00 or 11, is determined by whether the weighted sum $\sum_j w_j x_j$ is less than or greater than some *threshold value*. Just like the weights, the threshold is a real number which is a parameter of the neuron. To put it in more precise algebraic terms:

$$\text{Output} = \begin{cases} 0 \quad if \;\; \sum_j w_j x_j \;\leq\; threshold \\ 1\; if \;\sum_j w_j x_j \;>\; threshold \end{cases}$$

That's all there is to how a perceptron works.

VI. ENTROPY FACTOR

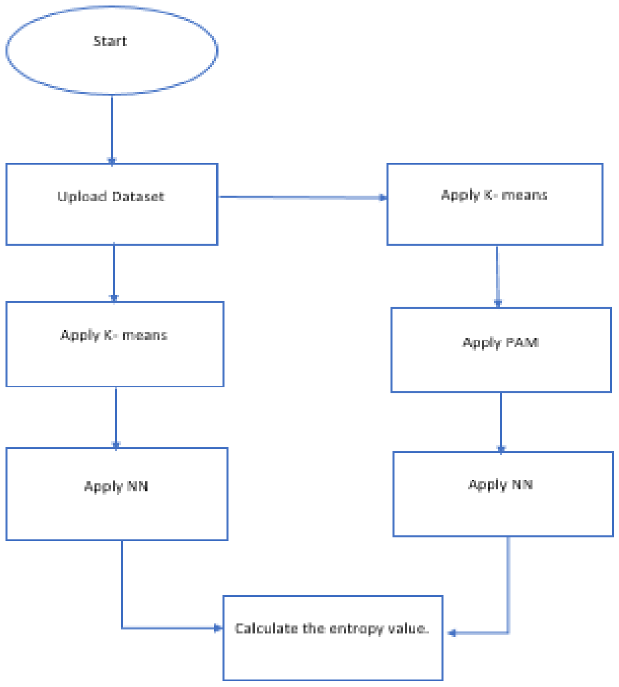
Supposing a discrete random variable X, which has x1, x2 ,..., xn , a total of n different values, the probability of xi appears in the sample is defined as P(xi), then the entropy of random[4] variable X is:

Hp = p (xi) logp(xi)

Entropy value ranges between 0 and 1. If H (P) = 0 (means close to 0), it indicates the lower level of uncertainty, and the higher similarity in the sample. On the other hand, if H (P) =1, it indicates the higher level of uncertainty, the lower similarity in the sample.

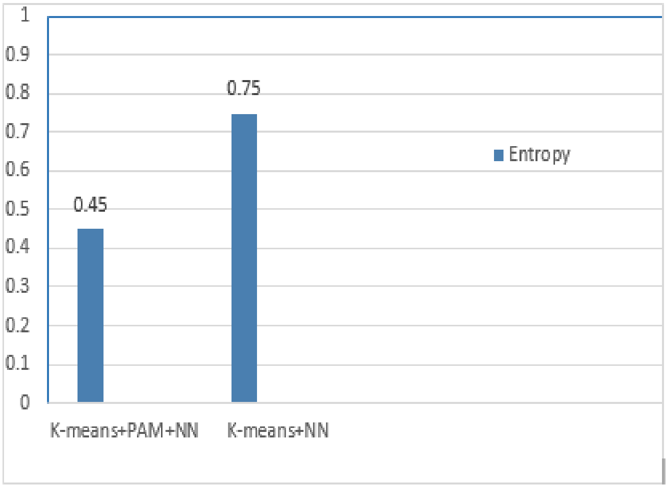
For instance, in the real network environment, for a particular type of network attack, the data packets show a certain kind of characteristics. For example, DoS attacks, the data [14] packets sent in a period of time are quite more similar in comparison to the normal network packets, which show smaller entropy, that is, the lower randomness.

VII. WORKFLOW



VIII. RESULT AND DISCUSSION

The whole implementation is done in NetBeans. The following table and graph shows the accuracy results of the proposed technique



Approach	Parameter	Value
K-means +NN	Entropy	.23 - .45
K-means+PAM+NN	Entropy	.50 - .75

IX. CONCLUSION

This paper proposes a new approach added in k-means in order to improve the efficiency and reduce the entropy. y. In order to label the unlabeled data, we have presented classification by NN because they can be effectively used for noisy data and it can also work on untrained data.

This approach can be expanded in various fields and This method ensures the entire process of clustering time will be reduced in execution process

REFERENCES

[1] Quan Qian, Tianhong Wang and Rui, Zhan, "Relative Network Entropy based clustering Algorithm for Intrusion detection", Vol.15, No. 1,pp.16-22, Jan,2013.

[2] Parampreet Kaur1 , Mr. Sahil Vashist2 , Roopkamal Ahluwalia3 , Gagangeet Singh Aujla4 "Entropy Reduction Based On K-Means Clustering And Neural Network/SVM Classifier" in International Journal Of Engineering And Computer Science Volume 3 Issue 11 November, 2014 Page No. 9166-9168

[3] AthmanBouguettaya , Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song "Efficient agglomerative hierarchical clustering" Elsevier 2014.

[4] PrasathPalanisamy, Perumal, K.Thangavel, R.Manavalan "A Novel Approach to Select Significant Genes of Leukemia Cancer Data Using K-Means Clustering" IEEE Pattern Recognition, Informatics and Mobile Engineering 2013.

[5] Palwinderkaur ,Usvirkaur ,Dr.Dheerendra Singh "Hybrid Clustering and Classification for Entropy Reduction: A Review" International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2014.

[6] Shalu Sharma1 , Sukhvinder Kaur2 , Ms. Jagdeep Kaur3 "Hybrid Clustering and Classification" International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 1, January 2015.