Volume 5, Issue 1, January 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hybrid Clustering and Classification

Shalu Sharma[1], Sukhvinder Kaur[2], Ms. Jagdeep Kaur[3]

[1]Research Scholar ,Department of CSE, PTU, India

[2]Assistant Professor (CSE),HPTU Hamirpur, KCIET Pandoga Una (H.P.), India

[3]Assistant Professor & HOD(CSE\IT), KCCEIT SBS Nagar, India

Abstract-In data mining, C-Means clustering is well known for its efficiency proved good for large data sets. The aim of every clustering algorithm is to group the similar data items while ungroup the dissimilar items. C-Means clustering algorithm has the opposite principle as fuzzy clustering algorithm has i.e. in C-Means every point has belonging to clusters while in fuzzy clustering, they belong to only one cluster.Clustering is a supervised learning algorithm. Clustering dispersion called entropy factor is the disorderness that occur after the clustering process. Less entropy leads to good clustering. Clustering with C-mean results in unlabeled data. I present a clustering algorithm called C-Means. Then unlabelled data is matched through neural classifier. Neural Network is the classification function to distinguish between members of the two classes in the training data. For classification we use Neural Network as they can recognize the patterns. The whole work is taken place in the Mat lab 7.10 environment in which entropy is taken as the main parameter for performance.

Keywords: Entropy, Clustering, Classification, Data Mining.

I. INTRODUCTION

Data mining is the process of automatically finding useful information in large data repositories [1]. Data mining rule called association rules are helpful in gaining similar items from database. As it is the process of grouping similar data. E.g. in supermarket we can predict which items are continuously demanding in the market by consumers. As per Wikipedia definition data mining consists of six steps:

- Anomaly detection
- Association rule learning
- Classification
- Clustering and
- Regression

The main purpose of the clustering algorithm is to find similar data from unlabeled data. The main purpose is finding a structure in a collection of unlabeled data. Clustering involves the grouping of the similar data.

Clustering algorithms can be loosely categorized into the following categories: hierarchical, partition-based, density-based, grid-based and model-based clustering algorithms [1-2, 3].

Among them, partition-based algorithms which partition objects with some membership matrices are most widely studied. Traditional partition-based clustering methods usually are deterministic clustering methods which usually obtain the specific group which objects belong to, *i.e.*, membership functions of these methods take on a value of 0 or 1.We can accurately know which group that the observation object pertains to. This characteristic brings about these clustering methods' common drawback, that we cannot clearly know the probability of the observation object being a part of different groups, which reduces the effectiveness of hard clustering methods in many real situations. For this purpose, fuzzy clustering methods which incorporate fuzzy set theory [4,5] have emerged. Fuzzy clustering methods [6-8] quantitatively determine the affinities of different objects with mathematical methods, described by a member function, to divide types objectively.

Among the fuzzy clustering method, the fuzzy c-means (FCM) algorithm [9] is the most well-known method because it has the advantage of robustness for ambiguity and maintains much more information than any hard clustering methods. The algorithm is an extension of the classical and the crisp k-means clustering method in fuzzy set domain. It is widely studied and applied in pattern recognition, image segmentation and image clustering. So in our proposed method we will use C Means Clustering algorithm.

Dividing objects in meaningful groups of objects or classes (cluster) based on common characteristic, play an important role in how people analyze and describe the world. For an example, even children can quickly label the object in a photograph, such as buildings, trees, people and so on. In the field of understanding data [10] we can say clusters are potential classes, and cluster analysis is a studying technique to find classes. Before discussing about clustering technique we need to provide a necessary description as a background for understanding the topic. First we define cluster analysis and the reason behind its difficulties, and also explain its relationship to other techniques that group data. Then explain two subjects, different ways of grouping a set of objects into a set of clusters and cluster types.

II. PROBLEM STATEMENT

In the searching of any particular data, clustering plays a vital role. If the cluster is not proper the searching may involve a lot of time and also the search results may not be exactly accurate. Hence the problem of this research is to enhance the clustering technique by combining the clustering algorithm with the classifier NEURAL CLASSIFIER. It would have to be seen that what algorithm of SVM would take the input of the c mean to produce an efficient cluster. The main aim of this research work is to reduce the entropy which the user gets after the searching.

The parameters of the evaluation would be as follows.

1. Time to retrieve the data
2. Entropy of the data
3. Cluster true and false positive

This occurs as due to dispersion in clustering. So we have to solve this problem using the Mat lab environment.

III. IMPLEMENTATION FRAMEWORK

Text clustering is a process of partitioning a set of text objects into clusters such that texts in the same cluster are more similar to each other than texts in different clusters according to some defined criteria. The main aim of text clustering is to find the structure feature of a set of text objects by partitioning it reasonably. And then we can [11] extract the hidden information from the texts using the clustering results. Note that the texts mentioned above contain news report, web, e-mail, paper,

Newsgroup articles etc [2, 3]. There are many methods that can be used to cluster texts. As a matter of fact, all kinds of clustering methods can be used to cluster texts in principle [4]. The commonly used clustering methods are: k-means, model estimation (especially mixed model estimation), hierarchical clustering (it can be classified into divisive and agglomerative types) etc. The FCM algorithm implements the clustering task for a data set by minimizing an objective-function subject to the probabilistic constraint that the summation of all the membership degrees of every data point to all clusters must be one. This constraint results in the problem of this membership assignment, that noises are treated the same as points which are close to the cluster centers. However, in reality, these points should be assigned very low or even zero membership in either cluster. In But this thesis proposes the C-Means method and NN approach for clustering method.

Step-1

1. Upload the data sample
2. Apply C-Mean to get the clusters on the basis of weight assigned to data
3. Classify the clusters using NN.
4. Optimize the entropy by getting support / confidence value.

Step-2

5. Upload the data sample
6. Then apply Neural Network (NN).
7. Repeat steps 3 and 4.
8. Finally get a comparison between C Mean / NN and SVM in clustering scheme on the basis of entropy metric.

- **Apply C mean**

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two [12,13] or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 , \qquad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centersc_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} , \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

This iteration will stop when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where $\varepsilon$is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

- **Apply NN**

A neural network is an interconnected group of nodes, a kin to the 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-

- Artificial method.
- Feed forward method

Advantages of using neural-

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

- **Apply entropy factor**

It is very important theory in the case of information theory (IT), which can be used to reflect the uncertainty of systems. From Shannon's [4] theory, that information is the eliminating or reducing of people understanding the uncertainty of things [4]. He calls the degree of uncertainty as entropy.

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. Another example is a network probing attack, which scans frequently a specific port in a certain period of time, so the destination ports will get smaller entropy compared with the random port selection of normal packets.

As an effective measure of uncertainty, the entropy, proposed by Shannon [5], has been a useful mechanism for characterizing the information content in various modes and applications in many diverse fields. In order to measure the uncertainty in rough sets, many researchers have applied the entropy to rough sets, and proposed different entropy models in rough sets. Rough entropy is an extend entropy to measure the uncertainty in rough sets. Given an information system *IS*

= (*U, A, V, f*), where *U* is a non-empty finite set of objects, *A* is a non-empty finite set of attributes. For any B⊆A, let *IND*(*B*) be the equivalence relation as the form of *U/IND*(*B*) = *{B*1,*B*2, ...,*Bm}*.

IV. MODEL STRUCTURE

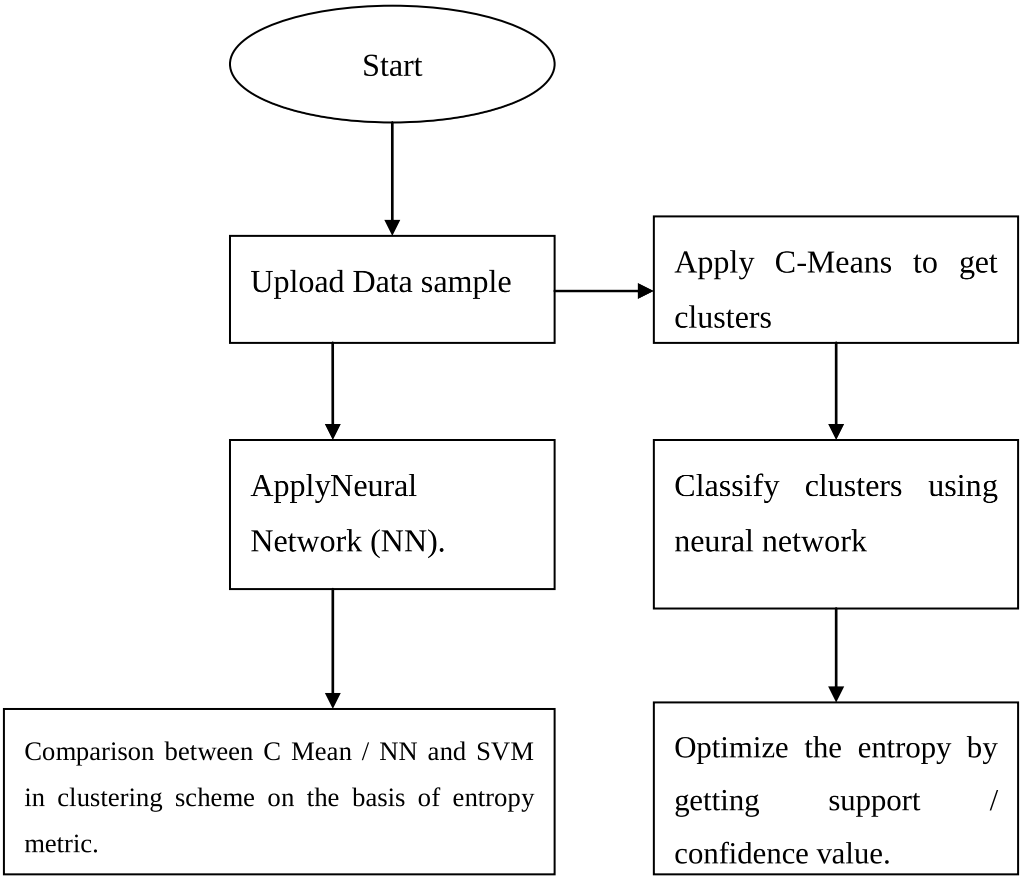
Start
Upload Data sample
Apply C-Means to get clusters
ApplyNeural Network (NN).
Classify clusters using neural network
Comparison between C Mean / NN and SVM in clustering scheme on the basis of entropy metric.
Optimize the entropy by getting support / confidence value.

Fig.1 Proposed Workflow

V. RESULT AND DISCUSSION

The whole implementation is taken place in MATLAB [15] environment. The following table and graph shows the accuracy results of the proposed technique.

mTable 1 Result table

Classifier	parameter	Value
C-mean	entropy	.23-.50
NN	entropy	.45-.70

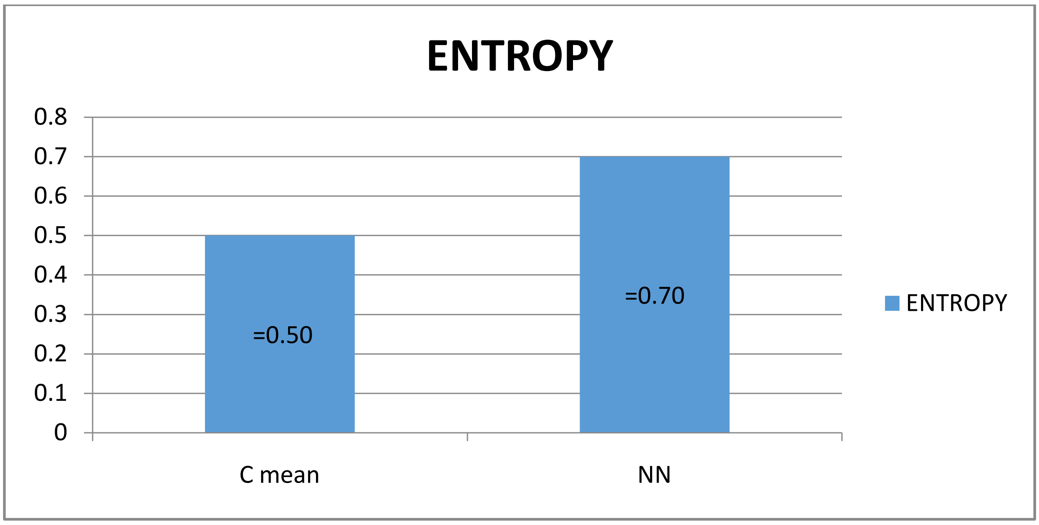

Fig .2 Comparison Figure

VI. CONCLUSION

In this thesis, we have presented C mean clustering algorithm as it is suitable for high dimensional data and also outlier detection occur efficiently. In order to label the unlabelled data, we have presented classification by NN because they can be effectively used for noisy data and it can also work on untrained data. Using this hybrid technique, entropy of the retrieved data can be reduced and also retrieval time, accuracy can be greatly enhanced.

In future, we can improve the traditional FCM algorithm in term of the selection strategy of initial cluster centres to fit the characteristics of meteorological data. The improved FCM algorithm has smaller squared errors than the traditional FCM algorithm while maintaining the rapid speed of convergence. Besides, the performance of the improved FCM algorithm is better than K-means algorithm in terms of the priori meteorological knowledge when implemented In Mat Lab [16]. We can improve the performance of the FCM algorithm in the field of meteorology from other aspects.

REFERENCES

[1] Hiroshi Motoda , Geoffrey J. McLachlan, Angus Ng, Bing Liu,Philip S. Yu ,Zhi-Hua Zhou ,Michael Steinbach , David J. Hand ,Dan Steinberg , “Top 10 algorithms in data mining” ,*knwlinfsyst*,Vol 14,pp.1-37,2008.

[2] Laura Auria and Rouslan A. Moro, “Support Vector Machines (SVM) as a Technique for Solvency Analysis”,2008.

[3] Feng Wen-ge , “Application of SVM classifier in IR target recognition” ,*Physics procedia*,Vol.24.pp. 2138-2142,2012.

[4] G. Karypis, E. H. Han and V. Kumar, J. Computer, vol. 32, no. 8, (1999).

[5] T. Zhang, R. Ramakrishnan and M. Livny, J. Data Mining KnowledgeDiscovery,vol. 1, no. 2,(1997).

[6] G. Sheikholeslami, S. Chatterjeeand A. Zhang, Editors,A. Gupta, O. Shmueli and J. Widom. Proceedings of the24th InternationalConferenceonVery Large Data Bases,New York, USA, (1998)August 24-27.

[7] W. Wang, J. Yangand R. Muntz, Editors. M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P.Loucopoulos and M. A. Jeusfeld,Proceedings of the23rd InternationalConferenceonVery Large Data Bases, Athens,Greece, (1997)August 25-29

[8] L. A. Zadeh, J. InfectionControl,vol. 8,(1965).

[10] http://www.tutorialspoint.com/data_mining/

[11] http://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd

[12] http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

[13] http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/ebooks/html/csa/node205.html

[14] http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[15] http://in.mathworks.com/products/matlab/

[16] http://in.mathworks.com/discovery/matlab-gui.html