www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242

Volume 3 Issue 11 November, 2014 Page No. 9166-9168

Entropy Reduction Based On K-Means Clustering And Neural Network/SVM Classifier

Parampreet Kaur[1] *, Mr. Sahil Vashist*[2]*, Roopkamal Ahluwalia*[3]*, Gagangeet Singh Aujla*[4]

Department of Computer Applications. CBS Landran,
Department of CSE, CEC Landran
(pkaler90@gmail.com), (sahilvashist90@gmail.com) , (roopkamal24@gmail.com),
(cecm.cse.gagangeet@gmail.com)

ABSTRACT: ***Clustering is the unsupervised learning problem. Better Clustering improves accuracy of search results and helps to reduce the retrieval time. Clustering dispersion known as entropy which is the disorderness that occur after retrieving search result. It can be reduced by combining clustering algorithm with the classifier. Clustering with weighted k-mean results in unlabelled data. This paper present a clustering algorithm called Minkowski Weighted K-Means. This algorithm automatically calculates feature weights for each cluster and uses the Minkowski metric (Lp) Unlabelled data can be labeled by using neural network and support vector machines. A neural network is an interconnected group of nodes, for classifying data whereas SVM is the classification function to distinguish between members of the two classes in the training data. For classification we use neural networks and SVM as they can recognize the patterns. The whole work is taken place in the Matlab.7 environment.***

KEYWORDS:Clustering, K Means, SVM, Neural Network, Entropy

.

I. INTRODUCTION

Data mining is the process of automatically finding useful information in large data repositories [1].The purpose of deploying data mining techniques is discovering important patterns from datasets and also provide capabilities to predict the outcome of a future observation such as market basket analysis, means that by using "*Association Rules"* learning the supermarket can determined which products are frequently bought together or to predict if the new customer will spend more than 100 $ for shopping today at the store. As for the Wikipedia definition, data mining involves six common tasks: Anomaly Detection, Association rule learning [2], Classification, Clustering and Regression. In this paper we discussed mostly on clustering class. Clustering is the most important unsupervised-learning problem as every problem is of this type. The main purpose is finding a structure in a collection of unlabelled data. Totally, the clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery. K-Means is arguably the most popular clustering algorithm; this is why it is of great interest to tackle its shortcomings. The drawback in the heart of this project is that this algorithm gives the same level of relevance to all the features in a dataset. This can have disastrous consequences when the features are taken from a database just because they are available. Another issue of our concern is that K-Means results are highly dependent on the initial centroids. To address the issue of unequal relevance of the features we use a clustering algorithm called Minkowski Weighted K-Means [3]. This algorithm automatically calculates feature weights for each cluster and uses the Minkowski metric (Lp) Unlabelled data can be labeled by using neural network and support vector machines.

Remaining paper is organized as: Section II shows the problem formulation, Section III shows the methodology steps, Section IV shows the proposed flowchart diagram, Section V show the results and discussion, Finally Section VI shows the conclusion part.

II. PROBLEM FORMULATION

Dividing objects in meaningful groups of objects or classes (cluster) based on common characteristic, play an important role in how people analyse and describe the world [4]. 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 we can say clusters are potential classes, and cluster analysis is a studying technique to find classes. [5] 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.

Web mining is the process of retrieving the data from the bulk amount of data present on web according to user need. This is important to the overall use of data mining for companies and their internet/ intranet based applications and information access. Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the [6] important information from customers visiting the site. But in the search results, there is often a lot of randomness and inaccuracy due to improper clustering and classification. Clustering is the unsupervised learning problem. Clusters are made on the basis of similar characteristics or similar features. Clustering is defined as the process to maximize the intercluster dissimilarity and minimize the intracluster dissimilarity. After [7,8] clustering, the classification process is performed so as to determine the labels for the data tuples that were unlabelled (no class). But Entropy is the disorderness that occurs after retrieving search results.

III. METHODOLOGY

Step-1

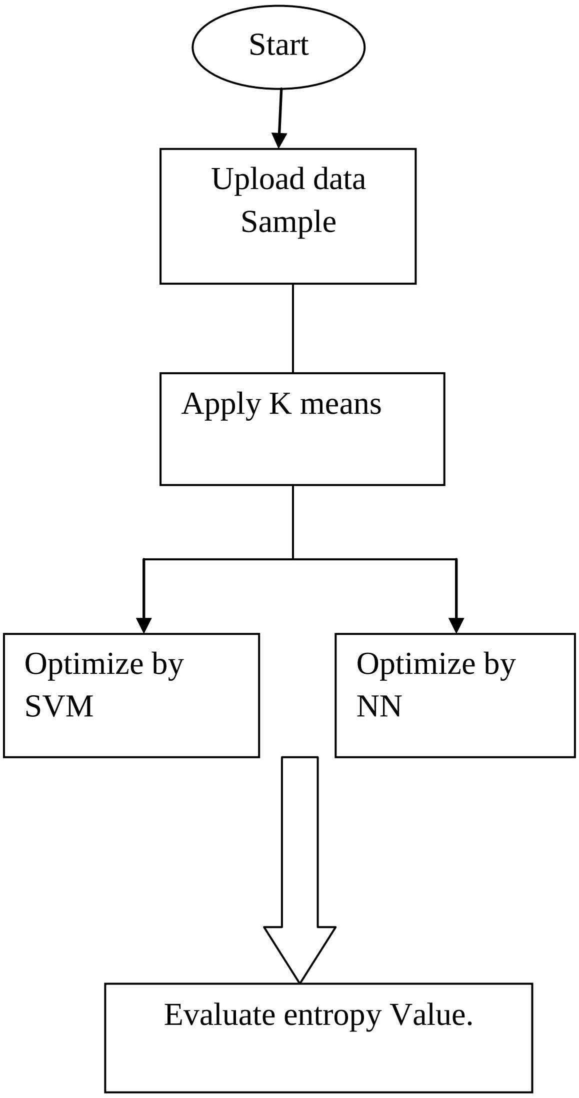
1. Upload the data sample
2. Apply Weighted k Mean to get the clusters on the basis of weight assigned to data
3. Classify the clusters using Neural Network.
4. Optimize the entropy by getting support / confidence value.

Step-2

5. Upload the data sample
6. Then apply Support Vector Machine (SVM)
7. Repeat steps 3 and 4.
8. Finally get a comparison between weighted K Mean and SVM in clustering scheme on the basis of entropy metric.

All of the above steps besides step are performed in Matlab environment and the system developed at the end is a Matlab software.

IV. PROPOSED FLOWCHART



V. RESULTS AND DISCUSSION

Result simulation has been taken place in MATLAB 7.10 environment. Entropy parameter has been taken to evaluate the performance. 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 theory, that information is the eliminating or reducing of people understanding the uncertainty of things. He calls the degree of uncertainty as entropy [9,10].

As an effective measure of uncertainty, the entropy, proposed by Shannon, 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

Table.1 Entropy Values

classifier	parameter	Value
W K Mean + NN	entropy	.23-.50
SVM + NN	entropy	.45-.70

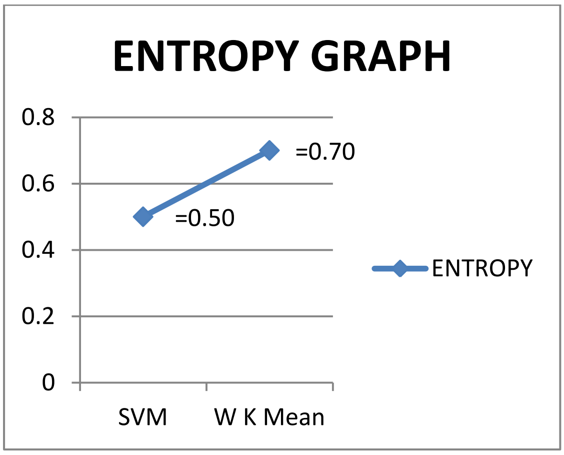


Figure 1 Comparison Figure of results

VI. CONCLUSION AND FUTURE SCOPE

In this work, we have presented weighted k 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 neural networks because neural can be effectively used for noisy data and it can also work on untrained data [11]. Using this hybrid technique, entropy of the retrieved data can be reduced and also retrieval time, accuracy can be greatly enhanced.

REFERENCES

[1] Vipin Kumar, Himadri Chauhan and Dhiraj Panwar, "K-Means Clustering Approach to Analyse NSL-KDD Intrusion Detection Dataset", Vol.3, Issue-4, Sept 2013.

[2]Liping Jing,Michael k.Ng,Joshua Zhexue Huang, " An entropy weighting k-mean algorithm for subspace clustering of high dimensional sparse data*", IEEE transactions on knowledge and data engineering,*Vol.19,no.8,August 2007.

[3] Son lam Phung and Abdesselam bouzerdoum, "A pyramidal Nueral network for Visual pattern recognition",*IEEE transactions on neural networks*,vol.18,no.2,March 2007.

[4]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.

[5]Xiangjun Li and Fen Rao "An rough entropy based approach to outlier detection", *Journal computational information systems*, Vol. 8 ,pp. 10501-10508, 2012.

[6] J. Y. Liang, Z. Z. Shi., "The information entropy, rough entropy, knowledge granulation in rough set theory", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12 (1), pp. 37 – 46, 2004.

[7]Velmurugan T., and Santhanam T., "Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithms for Statistical Distributions of Input Data Points," *European Journal of Scientific Research*, vol. 46, no. 3, pp. 320-330,2010.

[8] Z. Deng, K. Choi, F. Chung, and S. Wang, "Enhanced Soft Subspace Clustering Integrating Within-Cluster and Between Cluster Information," *Pattern Recognition*, vol. 43, no. 3, pp. 767-781, 2010.

[9] Xindong Wu, Vipin Kumar ,J. Ross Quinlan , Joydeep Ghosh , Qiang Yang, 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" ,*knwl inf syst*,Vol 14,pp.1-37,2008.

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

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