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

Entropy Reduction Using Hybrid Algorithm

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 - 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 unlabeled 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.

Keywords— Data Mining ,KDD, Clustering Mechanism , Neural Network .

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.

II. KDD PROCESS

Data mining or important part of Knowledge Discovery in Database (KDD), used to discover the most important information throughout the data, is a powerful new technology. Across a myriad variety of fields, data are being collected and of course, there is an urgent need to computational technology which is able to handle the challenges posed by these new types of data sets. The field of Data mining grows up in order to extract useful information from the rapidly growing volumes of data. It scours information within the data that queries and reports can't effectively reveal. As we mentioned earlier, the integral part of knowledge discovery in database (KDD) is data mining, which in our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. The KDD role is to convert raw data into suitable information as shown in figure 1.This process contains a series of transformation steps, from data pre-processing to data mining results.

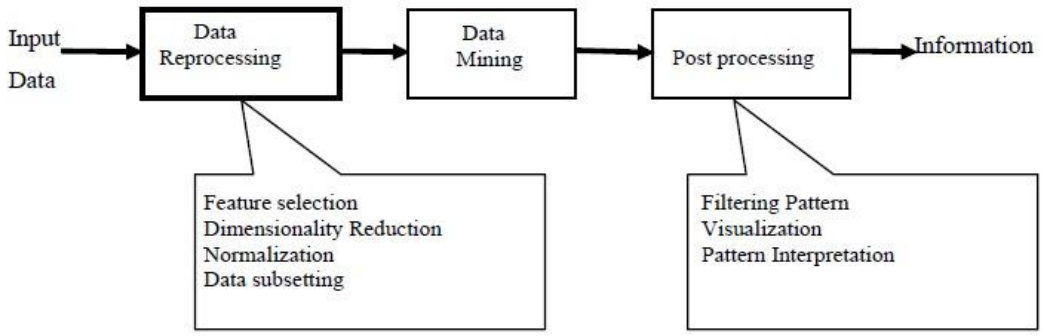

Fig.1 KDD Process

III. CLUSTERING MECHANISM

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.

IV. CLASSIFICATION ALGORITHM

In this paper, we propose an effective and novel scheme about the classification algorithm i.e neural network with the classifier c-mean clustering algorithm. To fully ensure the entropy is reduced while cluster or data set are properly managed. Neural networks are composed of simple elements which operate parallel. A neural network can be trained to perform a particular function by adjusting the values of the weights between elements. Network function is determined by the connections between elements. There are activation functions used to produce relevant output.

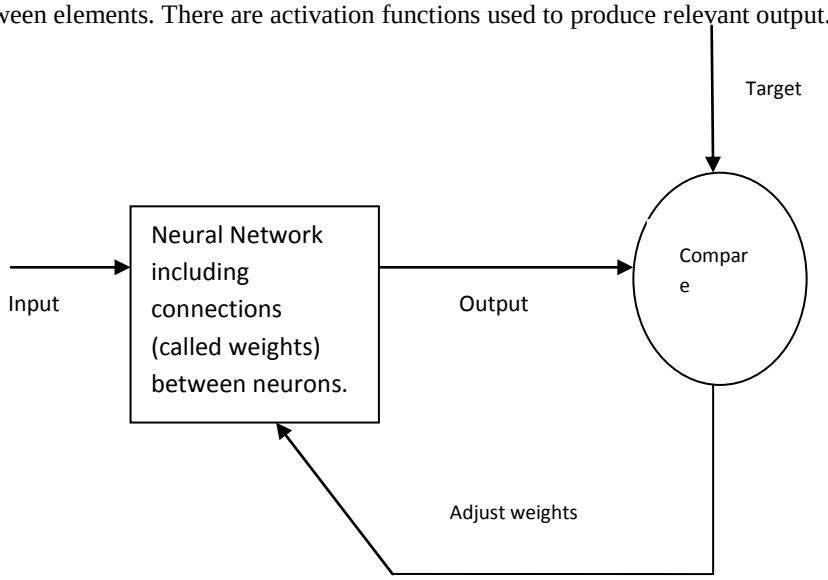

Fig. 2 The architecture of neural network

V. OBJECTIVES

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.

a) Time to retrieve the data
b) Entropy of the data
c) Cluster true and false positive

VI. METHODOLOGY

1. Apply C-Mean to get the clusters on the basis of weight assigned to data
2. Classify the clusters using NN.
3. Optimize the entropy by getting support / confidence value.
4. Then apply Neural Network (NN).
5. Finally get a comparison between C Mean / NN and SVM in clustering scheme on the basis of entropy metric

VII. RELATED WORK

In our literature survey, we note that DDoS defense mechanisms are generally classified as preventive mechanisms and reactive Mechanisms. [5] This paper describes their experiments, ranking methods with different supervised learning algorithms give quite different results for balanced accuracy and confirmed that in order to be sure that a subset of features giving the highest accuracy has been selected; the use of many different indices is recommended.

[4] .The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today's business environment, efficiency or speed is not the only key for competitiveness. This type of huge amount of data's is available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. This paper imparts more number of applications of the data mining and also o focuses scope of the data mining which will helpful in the further research.

[11] This paper represent the feature selection plays an important role in data mining field. The feature selection for continuous attributes is a hot issue in recent years. Firstly, combining the existing research and introduces the concept of entropy breakpoint into the discrimination of continuous attributes. Secondly, according to the defect which information gain tended to attribute with more values, the paper use the standardized gain to replace the information gain to measure the feature selection, and propose an algorithm for continuous attributes based on the information entropy feature selection. The experimental results show that, the algorithm has a better effect on high dimension data set.

VIII. CONCLUSIONS

In this paper we explained about methdology how these methods are to be achevied .work is under progress to entropy reduction method using different algorithm. But we observe that 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 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.

REFERENCES

[1] Agrawal, T. Imielinski, and A.Swami. *"Mining association rules between sets of items in large databases".* in the Proc. of the ACM SIGMOD International conference on Management of Data, 1993.

[2] Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., *"A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data",* volume 5, ISSN 1307-6884," Proceedings of World Academy of Science, Engineering and Technology, April 2005.

[3]. Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., *"A Generic Data mining System basic design and implementation guidelines",* A Technical Project Report of CYCYTprojectofSpanishGovernment.1998.WebSite: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935.

[4] Chander Sahu ,Dr. B.S Patel*"Application of Fuzzy ID3 To Forcast Seasonal Runoff"* International journal of Computer Technology And Electronics Engineering (IJCTEE) Volume 1,Issue 3.

[5] Danyang Cao, Nan Ma *"A Feature Selection Algorithm For Countinuous Attributes Based on The Information System"* 8:4 (2012)1467-1475

[6] Domingos, P. and Pazzani, M, *"On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning",* 29, 103-130 (1997) .

[7] Dunham, M. H., Sridhar S., *"Data Mining: Introductory and Advanced Topics",* Pearson Education,New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[8] Dhanalakshmi. P.; Palanivel. S.; and Ramaligam. V.,*"Classification of audio signals using SVM and RBFNN, In Elsevier, Expert systems with application"*, Vol. 36, pp. 6069–6075 (2008),.

[9] Darin Brezeale and Diane J. cook, *"Automatic video classification: A Survey of the literature"*, IEEE Transactions on systems, man, and cybernetics-part c: application and reviews, vol. 38, no. 3, pp. 416-430 Fellow. IEEE (2008).

[10] Freund, Yoav, and Schapire, R. E., *"Experiments with a new boosting algorithm. In Machine Learning":* Proceedings of the Thirteen International Conferences, pp. 148-156 (1996).

[11] Jen-Da Shie · Shyi-Ming Chen, *"Feature subset selection based on fuzzy entropy measures for handling classification problems"* , Appl Intell (2008) 28: 69–82,DOI 10.1007/s10489-007-0042-6.

