International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org
Vol. 8 Issue 1, January-2018,
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196 |

EXPLORING DIFFERENT DATA MINING TECHNIQUES FOR PROTEIN AND FOREST FIRE DATA

VRATNAKUMARI[1], DR. ARVIND KUMAR SHARMA[2]

Department of Computer Science and Engineering

[1,2]OPJS University, Churu (Rajasthan) – India

ABSTRACT

In recent years, data mining has been worn in many areas of knowledge and industrial, such as bioinformatics, medicine, health care, physics and environment. In the study of human genetics, genetics, sequence mining helps tackle the very important goal of accepting. Data mining idea and physical model from data has open for centuries. Early methods of identifying patterns in data comprise regression analysis and Bayes' theorem. The proliferation, ubiquity and growing power of computer technology has improved strategy gift, thought, data group and storeroom. Data mining concepts build up more facility for study. As data sets have adult in size has more and more been increased with indirect, mechanical data processing, aided by other discoveries in information technology and computer science such as neural networks, cluster analysis, genetic algorithms.

1. INTRODUCTION

1.1 Distributed data mining

Data mining algorithms deal predominantly with simple data formats (typically flat files); there is an increasing amount of focus on mining complex and advanced data types such as object-oriented, spatial and temporal data. Another aspect of this growth and evolution of data mining systems is the move from stand-alone systems using centralized and local computational resources towards supporting increasing levels of distribution [2]. As data mining technology matures and moves from a theoretical domain to the practitioner's arena there is an emerging realization that distribution is very much a factor that needs to be accounted for. Databases in today's information age are inherently distributed. Organizations that operate in global markets need to perform data mining on distributed data sources (homogeneous / heterogeneous) and require cohesive and integrated knowledge from this data. Such organizational environments are characterized by a geographical separation of users from the data sources [3]. This inherent distribution of data sources and large volumes of data involved inevitably leads to exorbitant communications costs. Therefore, it is evident that traditional data mining model involving the co-location of users,

data and computational resources is inadequate when dealing with distributed environments. The development of data mining along this dimension has lead to the emergence of distributed data mining. The need to address specific issues associated with the application of data mining in distributed computing environments is the primary objective of distributed data mining. Broadly, data mining environments consist of users, data, hardware and the mining software (this includes both the mining algorithms and any other associated programs) [4]. Distributed data mining addresses the impact of distribution of users, software and computational resources on the data mining process. There is general consensus that distributed data mining is the process of mining data that has been partitioned into one or more physically/geographically distributed subsets.

2. ALGORITHMS

2.1 K-means algorithm

We are using k-means algorithm in clustering K-means clustering is a partitioning method. K-means treats explanation in your data as matter having locations and distance from each other.

2.2 Fuzzy c-means (FCM)

Fuzzy C means algorithm is a method of clustering technique. FCM allows one piece of data to belong to two or more clusters. It's method perfection on earlier clusters method. Its method explored by Jim Bezdek in 1981 [5].

2.3 Bagging

Bagging works as a method of increasing accuracy. Bagging was produced by Breiman and it is also called Bootstrap aggregating and Bootstrap is based on random sampling with replacement. Bagging is a machine learning ensemble algorithm [6].

2.4 Stacking

Stacking is also called stacked generalization. Stacking involves training a learning algorithm. Stacking is combining to the predictions of several other learning algorithms. It is complicated. It performance is good other single one of the trained model [7].

2.5 Random Subspace algorithm

This research we also include Random Subspace. RSM is the combining technique. It is introduced by TK. An ensemble is a collection of base learners. Ensemble is also called base classifier. RSM described in WEKA tool [8].

2.6 Extra work

Clustering technique is dividing data elements into different groups. These groups are called as clusters [9]. The elements within a group possess high

similarity while they differ from the elements in a different group.

2.7 Gaussian mixture model-

Gaussian mixture model is multivariate distribution. It is a mixture of one or more multivariate Gaussian distribution component. Gaussian component is defined by a vector of mixing proportions in Gaussian mixture model for each multivariate distribution [10].

2.8 Objective

The objective of the present research is to explore different data mining techniques for the protein, and forest fire data.

3. METHODS OF THE RESEARCH

3.1 Tools

In this research we are using **MATLAB** (**mat**rix **lab**oratory) R2007b. It's is a numerical computing surroundings. MAT LAB allows smanoeuvrings of functions,algorithms, matrix manipulation s creation of user interfaces, and interfacing with programs written in other languages. MATLAB based on fourth generation programming language. MATLAB is involving C, C++, Java, and FORTRAN.

We use WEKA manual version 3-6-9 tool. WEKA says that it is a decision tree based classifier which constructs multiple trees in randomly selected feature subspaces. However, WEKA allows the base classifier for the ensemble to be any algorithm (not necessarily trees). These are called Classifiers in WEKA. We use forest fire data set in WEKA with classification algorithm.

4. FCM ALGORITHM USING FUZZY LOGIC TOOLBOX AND K-MEANS ALGORITHM USING

4.1 MATLAB on protein data set

In the ancient times several types of data breakdown task and various types of data in environment tended to handle rather small data sets. Very huge quantity of data has been cool from genetic knowledge remedial, strength care and in surroundings. The database is increased even quicker and starts more density and hitches in data. Data Mining is a awareness detection from Data.

Data mining has evolved and continues to evolve from the fork of research fields such as machine education, natural science, healthcare, industry, pattern appreciation databases, and statistics and with the help of data mining confiscate complication in data. In this research we argue MEROPS online tool for protein data set. Protein sequences those are nearer to each other. So with the assist of data mining method we at hand how can take away this problem and continue the reserve in dataset with the help of clustering. We at hand this

research MEROPS online tool for protein data set, FCM algorithm, Fuzzy logic toolbox, K-means algorithm, MATLAB, bioinformatics toolbox in MATLAB.

4.2 Implementation of k-means algorithm

We are implementing K-Means algorithm in MATLAB software**.** Performance measures in this algorithm such as No. of iterations, Create Clusters and establish division, Determine the exact Number of Clusters, Avoid Local Minima, Silhouette validity index Accuracy. K-means first loads data idx = k-means (data, n*),* which partitions the direct in the n-by-p statistics medium statistics into n clusters. We are analyzing the result in different clusters.

4.3 Comparative Analysis of bagging, stacking and random subspace methods on forest fire dataset using WEKA tool

Data mining is a powerful new technology and its tools predict behaviors and future trends. Data mining techniques is very important in the analysis of real environmental data. Forest fire is important to the forest ecosystem. Only few researchers work on scientific data. Analyzing the scientific data is very different task. This research includes Data Mining concepts and methods. We present a Data Mining technique to analyze and improve accuracy of the forest fires data set and comparative study is done between Bagging, Stacking and Random subspace algorithms into WEKA tool. In our research we introduce how data mining technique works. We present an integration of the algorithm on analysis forest fire data set in to WEKA data mining suite and we compare better results of these methods. We are using bagging, stacking and random subspace algorithms with two classifiers. First classifier is decision stump and second classifier is decision table. Decision stump is a machine learning model. Decision stump is also called single rules.

Decision table like a decision trees algorithm. Presentation marks explain that the classifiers built. These classifiers are more accurate than that produced by the classification methods. Many researchers used different methods in this experiment. We are using Bagging, random subspace and stacking. Bagging works as a method of increasing accuracy of classifier models. It's generating random. Its independent bootstrap replicates. Bagging is example of ensemble methods. Stacking is called learning algorithm and it is used to combine the predictions. Random Subspace ensembles (RSE) base classifiers as the highly accurateclassification method. Random Subspace Method (RSM) says that it is a decision tree based classifier in WEKA.

Bagging and Random subspace methods have become popular combining

techniques. These methods are useful for improving weak classifiers and improving accuracy. Decision tree based classifier is useful highest accuracy on training data. Decision Table algorithm classifier is useful to summarizes the dataset. It is useful to find subset of attributes. We are taking the decision table algorithm with bagging, stacking and random subspace algorithm. In this research we are using regression analysis in WEKA tool. Regression analysis is a statistical tool. It is a relationship between variables

5. ANALYSIS

In experiment we take an intuitive look at how increasing accuracy. In all experiments evaluated are using 20-fold cross validation. Performance result is depends on the classification error because data set isregression task. In regression task lowest mean error of data set predict the higher accuracy.

We are using bagging, stacking and random subspace algorithms with two classifiers. First classifier is decision stump and second classifier is decision table. Decision table like a decision trees algorithm. The experimental results are created using WEKA 3-6-2, open source software.We are these 3 algorithms using with MATLAB software.

First is k-means algorithm second is k-means algorithm with Gaussian mixture model and third is fuzzy c means algorithm. Firstly, we are taking data points through k-means algorithm and allows k-means algorithm. Secondly, we are taking data points through Gaussian mixture model and allows k-means algorithm also lastly, we are taking data point through fuzzy c means algorithm and allows fuzzy c means algorithm. The performance evaluated on the forest fire dataset. We are taking data set from UCI machine learning. Firstly we have compared k-means algorithm and k-means algorithm with Gaussian mixture model.

6. RESULTS

6.1 Result and discussion in MATLAB

In experiments, In table 1and 2 results presented, classification errors are using 20-fold cross validation the performance results of stacking algorithm able to improve accuracy. So we can say that in terms of performance, stacking with decision stump learning tree and decision table algorithm both gives better results than other methods. The algorithms are implemented using MATLAB and fuzzy logic tool box and results are evaluated based on performance parameter in both algorithms. After doing this research experiment results show that how k-means and fuzzy C means implemented on protein data set. In this research work we removed the problem that show proteins are highly affiliated to each other.FCM allows one piece of

data to belong to two or more clusters. We used different clusters in our research. Results based on different clusters in both algorithms. K-means is the centroid based technique. We are also compared k-means and FCM results in this research. Comparison results show that the k-means is better than FCM.Bagging, stacking and random subspace algorithms are implemented using WEKA and experiments are conducted and results are evaluated based on performance. Finally, we compared performance of bagging, stacking and random subspace algorithm. On the basis of experiments, we have found that using 20-fold cross validation then performance of the stacking with decision stump and decision table improve the prediction accuracy of classifier. So stacking is better and straightforward to interpret other. Stacking algorithm built accurate classifier model and consuming less time.When we used data points through k-means algorithm then k-means gives better solutions, but when we used data points through fuzzy c means then FCM algorithm is not able to improve better performance. This research is useful to found the quality of k-means algorithm. This algorithm suitable for increases the accuracy in future. K-Means will also be modified using the other methods in future.

The finding comparison result in between MATLAB and WEKA tool explore that the accuracy rate is high in MATLAB as compared to WEKA tool.

Table 1: Bagging, Stacking, and Random subspace algorithm with Decision Stump result on forest fire data set

Comparison field	Bagging with DS	Stacking with DS	Random subspace with DS
Mean absolute error	19.05	16.33	18.51
Root mean squared error	41.97	39.42	40.64
Relative absolute error	102.95	84.23	99.70
Root relative squared error	112.40	94.41	104.62
Time	0.04	0.02	0.03

Table 4.17: Evaluation of bagging, stacking and random subspace algorithm with decision table on forest fire data set

Comparison field	Bagging with DT	Stacking with DT	Random subspace with DT
Mean absolute error	18.56	18.57	18.60
Root mean squared error	40.34	40.05	40.10
Relative absolute error	99.63	99.92	100.15

Root relative squared error	101.30	99.86	100.29
Time	5.4	0.02	2.37

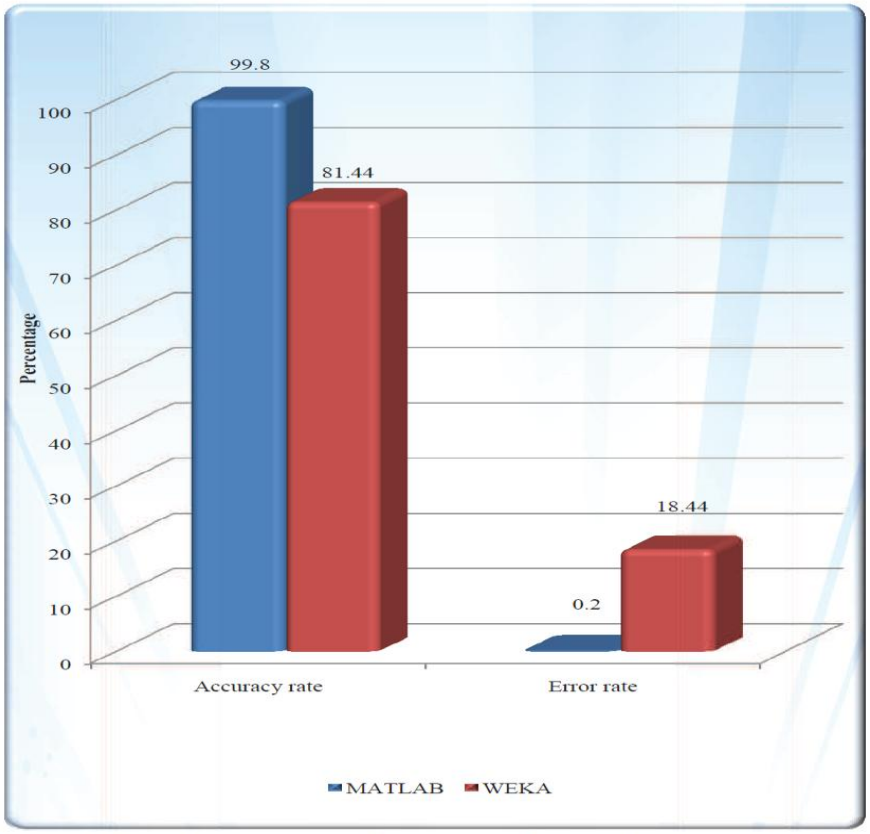

Figure 4.23: Graph representing the comparison between the accuracyrate and error rate of WEKA and MATLAB.

7. CONCLUSION

Data mining is very useful technology in future. This research is useful for analysis typical data. Researcher can easily remove complexity from data sets in future. K-means algorithm may be suitable for clusters are “similar” to one another. Clusters are “dissimilar” to object in other clusters and it’s may found the similarity and dissimilarity in between the other data set.In environment many types of typical data set is found. Only a fewer researcher explored the scientific data. WEKA tool research is useful for uses predictive Classifier models of forest fire data set and can be analyzing complexity of data set.In near future we will extend this work by determines the othermachine learning algorithm to enhance the accuracy rate further. K-Means may also be modified using the other appropriate methods.

REFERENCES

[1] Yu Poh Yong, Omar Rosli, Harrison D. Rhett, Sammathuria Kumar Mohan, Nik Rahim Abdul (2011), "Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods", Journal of Computational Biology and Bioinformatics Research 3(4): 47-52.

[2] Zeng, L., Li, L., Duan, L., Lu, K., Shi, Z., Wang, M., Wu, W., Luo, P. (2012), "Distributed data mining: a survey", Information Technology and Management, 13(4): 403-409.

[3] Verbeek, J.J., Vlassis, N., Krose, B. (2003), "Efficient Greedy Learning of

Gaussian Mixture Models", Published in Neural Computation, 15(2): 469-485.

[4] Trees", Proc European Conference on Machine Learning, 9 :323-334.

[5] Smith Dave, SAS, Marlow, UK (PUSH 2007) Data Mining in the Clinical Research Environment.

[6] Raut A. B, Nichat A. A. 2017. Students Performance Prediction Using Decision Tree Technique International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13 pp. 1735-1741

[7] Poddar, Sudip and Mukhopadhayay, Anirba (2012), "Cluster: A MATLAB GUI Package For Data Clustering. International Technology Research Letters", 1 (1): 33-48.

[8] Mrs. Sujatha B. and Akila C., (2012), "Prediction of Breast Cancer Using Data Mining", International Journal of Computer Science and Management Research, 1 (3): 384-389.

[9] KaurSumit, Bansal R. K. 2016. Mixed Pixel Clustering and Classification Techniques: A Review *International Journal on Recent and Innovation Trends in Computing and Communication* Volume: 2 Issue: 5 1054 – 1059

[10] Efron B. and Tibshirani R., (1993) "An Introduction to the Bootstrap", Chapman & Hall, New York.