Prediction Analysis Techniques of Data Mining: a Review

[1]Adeep Malmotra, [2]Bhavna Arora
[1]M.Tech Student, [2] Assistant Professor
[1]Department of Computer Science & IT, Central University, Jammu,
[2]Department of Computer Science & IT, Central University, Jammu
[1]adeepmalmotra30@gmail.com, [2]bhavna.aroramakin@gmail.com

ABSTRACT

The technique through which important information is extracted from the raw data in data sets is known as data mining. The future scenarios related to current data can be predicted with the help of prediction analysis technique provided under data mining. Clustering and classification forms the basis of prediction analysis. Numerous techniques have been proposed by various researchers in order to perform prediction analysis on various real-time applications. This paper describes the various techniques of prediction analysis proposed by various researchers. The paper also presents a review and analysis of these techniques based on parameters such as algorithms and techniques, datasets, attributes and tools used for analysis.

Keywords: - Prediction analysis, Classification, Clustering, K-means, SVM (Support Vector Machine)

INTRODUCTION

Data mining is the process of extraction of interesting patterns and knowledge by analysing data. Various data mining tools are available which can be used to analyse different types of data (varying from structured to unstructured data). Decision making, market basket analysis, production control, customer retention, scientific discoveries and education systems are some of the application areas that uses data mining for analyzing the collected data [1] to yield useful results. Classification and Clustering are the two techniques which are used extensively for mining important information from the data. Data is either classified or clustered depending upon the availability of the training data. If training data is present, supervised methods of classification such as Naïve Based, SVM (Support Vector Machine), and Regression etc. are used. In the absence of the training of training data, unsupervised clustering algorithms such as k-means clustering, hierarchical clustering and k-medians etc. are used.

A. Clustering in Data Mining

Clustering is the process of organising data into clusters based on their similarity index i.e. the data elements in one cluster are more similar than the elements in other clusters. The clusters are generated by analysing similar patterns of the input data. The unsupervised data clustering

classification method create clusters, group of objects in such a way that objects in different clusters are distinct and that are in same cluster are very similar to each other. In data mining, cluster analysis is considered as one of the traditional topic which is applied for the knowledge discovery. The data objects are grouped into a set of disjoint classes which is known as clusters [2]. Objects within a class have high resemblance to each other and these objects divided into separate classes are more distinct. For example in biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in population. In a city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used to classify all documents available on Web. Following are some broader categories into which the clustering methods have been categorized:

- **Partitioning Methods**:-The gathering of samples that are of high similarity in order to generate clusters of similar objects is the basic functioning of this method. Here, the samples that are dissimilar are grouped under different clusters from similar ones. These methods completely rely on the distance of the samples [3].
- **Hierarchical Methods**:-A given dataset of objects are decomposed hierarchically within this technique. There are two types in which this method is classified on the basis of type of decomposition involved. They are agglomerative and divisive based methods [4]. A bottom up technique in which the formation of separate group is the first step performed is known as agglomerative technique. Further, the groups that are near to each other are merged together.
- **Density Based Methods:-**The distance amongst the objects is taken as a base in order to separate the objects into clusters in most of the techniques. However, in this methods clusters of arbitrary shapes are identified by distinguishing low density regions from high density regions. The concept of density connectivity and density reachability are used to find density areas or regions.
- **Grid Based Methods**:- A grid structure is generated by quantizing the object space into finite number of cells which is known as grid based method. This method has high speed and does not depend on the number of data objects available.

B. Classification in Data Mining

Classification is the process of categorizing the given data into certain defined sets of outcomes. The output category is already defined by the user and with the help of the techniques like regression, svm, etc., the given data is classified into one of the categories as defined by the user. Consider an example, of analysing the current weather information of a particular day and classifying them into categories such as “sunny”, “rainy “ or “cloudy.

Two steps are followed within this process. They are [5]:

- **Model Construction**: Model construction describes the set of predetermined classes. The class label attribute determines each tuple/sample which is assumed to belong to a

predefined class. Wide numbers of tuples are used for the construction of the model known as training set. They are represented as classification rules, decision trees, or mathematical formulae/regression.

- **Model usage***:* The second step used in the classification is model usage. In order to classify the test data, the training set is designed from the unknown data for the accuracy analysis. The classified result from the model is used to compare with the known label of test sample. Test set is not dependent on training set.

This paper is organized in five sections. Section 1, gives the introduction of various clustering and classification techniques which are widely used in predictive analysis. Section 2, describes the details about the SVM classifier. The related work in prediction analysis is described in Section 3. A comparative analysis of the techniques used by researchers is presented in Section 4. Finally the paper concludes in Section 5.

SVM CLASSIFIER

SVM (Support Vector Machine) classifier was proposed for regression, classification as well as general pattern recognition. This classifier is considered to be good in comparisons to other classifiers because of its high generalization performance which does not require any prior knowledge. The performance is even better when the dimension of the input space is extremely high. In order to differentiate between the two classes of the training data, the SVM requires identifying the best classification function. The best classification function metric can be represented geometrically by means of hyperplane and margins [7]. The amount of space or distance amongst two classes is known as hyperplane. The shortest distance between the closest data point to a point on the hyperplane is known as margin. The hyperplane *f(x)* is separated through the linear classification function for the linearly separable dataset (as shown in the Fig.1). This hyperplane passes through the middle of two classes which separates them. The new data instance x_n is classified by testing the sign function $f(x_n)$; x_n which belongs to the positive class if $f(x_n) > 0$. This is done after the determination of a new function.

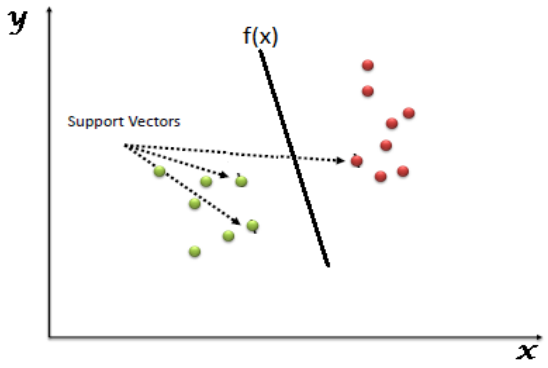

Fig. 1. SVM Classifier

The main objectives of SVM classier are defined as under:

- To determine the best function by maximizing the margin between the two classes. This is due to the fact that there are many such linear hyperplanes. This can help us in extending the margin by selecting only a few hyperplanes for the solution to SVM even when so many hyperplanes are available [8].
- To produce linear function which can help in identifying the target function. This helps in extending the SVM for performing regression analysis. The error models are helpful for the SVRs (Support Vector Regression). In case when the differences between the actual and predicted values are within an epsilon amount, the error is to be defined as zero. In the off chance, there is a linear growth in the epsilon insensitive error. Through the reduction of Lagrangian multipliers, the support vectors can be studied. The insensitivity to the outliers can be of benefit for the support vector regression.

The demerit of SVM is that the computations are not efficient enough. There are many solutions proposed for this. The breakage of one big problem into numerous numbers of smaller problems is one way to solve this issue. There are only some selected variables for the efficient optimization of each problem. Until all the problems are solved eventually, this process keeps working in iterative nature. The problem of learning SVM is to be solved also by recognizing the approximate minimum enclosing ball for the set of instances in the program.

LITERATURE REVIEW

In paper [9], the author has proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. Data was gathered from a hospital which includes both structured as well as unstructured data. In order to make predictions related to the chronic disease that had been spread in several regions, various machine learning algorithms were streamlined here. 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

In paper [10], the author has proposed a foggy k-mean clustering approach and tested it on a real time lung cancer dataset taken from SGPGI(Lucknow). The results of foggy k-mean and k-mean clustering algorithms are compared on 2 and 3 cluster respectively. Foggy k-mean clustering algorithm performs better for all validation measures such as Dunn index, connectivity and silhouette.

The author has presented weather forecasting analysis using clustering in paper [11]. A generic methodology has been used for Weather forecasting with the help of proposed incremental K-mean clustering algorithm. The weather events forecasting and prediction becomes easy using modeled computations. The author has performed different experiments to check the proposed approach for correctness.

Student Performance Analysis System (SPAS) for keeping the track of student's result was proposed by author in paper [12]. Student's performance is predicted by means of data mining techniques based on the grades scored by the students in a particular subject. Different classification techniques are compared on training data and then BFTree is chosen for training the data.

Stock market prediction software for determining the right time of buying and selling the stocks is developed by the author in paper [13]. The past historical knowledge of experiments has been used by stock market investors to predict better timing for buying or selling stocks. There are different data mining techniques available out of which decision tree classifier has been used by authors in this work.

The author in paper [14] has presented a study related to disease classification in medical field. In this field, every single day a large amount of data is generated and it is difficult to handle this much large amount of data. The K-means algorithm has been used to analyse different types of diseases. The cost and human effects has been reduced using proposed prediction system based data mining.

The real and artificial datasets for heart diseases are examined using the k-mean clustering technique by the author in paper [15] and used the results to predict the diagnosis of heart diseases to check its accuracy. The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis and each cluster has its observations with nearest mean. The proposed scheme of integration of clustering has been tested and its results shows that highest robustness, accuracy rate can be achieved by using it.

An improvised k-mean clustering algorithm has been proposed by the author in paper [16]. Data containing similar objects has been divided into clusters. A data of similar objects are in same group and in case dissimilar objects occur then it will be compared with objects of other groups. The proposed algorithm has been tested and results shows that the algorithm is able to reduce efforts of numerical calculation, complexity along with maintaining an easiness of its implementation. The proposed algorithm is also able to solve dead unit problem.

COMPARITIVE ANALYSIS

A comparative analysis of the various techniques used by the researchers on the basis of techniques/algorithm used, dataset on which the algorithm is applied, number of attributes taken by each, tools used for analysis and results is presented in Table 1

Research Reference	Techniques / Algorithms	Datasets used	No of Attributes taken for analysis	Tools Used for Analysis	Results
Min Chen[9]	Naïve Bayesian, KNN and Decision tree	Heart Diseases	79	MATLAB	Decision tree performs better as compared to others.
Akhilesh Kumar Yadav,[10]	Foggy K-mean Algorithm	Lung cancer Data	9	WEKA	Foggy k-mean performs well as compared to K-means
Sanjay Chakrabotry,[11]	Incremental k-mean clustering Algorithm	Air pollution Data	7	WEKA	The accuracy of proposed method is achieved up to 83.3 percent
Chew Li S.,[12]	BF Tree classifier	Student Performance	9	WEKA	BF Tree performs well as compared to other tree classifiers
Qasem A.,[13]	Decision tree	STOCK Data Prediction	170	WEKA	C4.5 classifier perform well as compared to ID3

Table1. Comparative Analysis of Prediction Analysis

CONCLUSION

The prediction analysis is the technique of data mining which is used to predict future from the current data. The prediction analysis is the combination of clustering and classification. The clustering algorithm groups the data according to their similarity and classification algorithm can assign class to the data. In this paper, various prediction analysis algorithms are reviewed and analyzed in terms of various parameters. The literature survey is done on the various techniques of prediction analysis and a comparative analysis of these is also done in this paper.

REFERENCES

[1] Abdelghani Bellaachia and Erhan Guven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.

[2] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, 2010, pp. 123-128.

[3] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.

[4] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, vol. 7, 2012, pp-56-62.

[5] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, vol. 5, 2012, pp. 959-963

[6] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", 2009, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol , IMECS.

[7] Chuan-Yu Chang, Chuan-Wang Chang, Yu-Meng Lin, (2012) "Application of Support Vector Machine for Emotion Classification", 2012 Sixth International Conference on Genetic and Evolutionary Computing, volume 12, issue 5, pp- 103-111

[8] Himani Bhavsar, Mahesh H. Panchal, (2012) "*A* Review on Support Vector Machine for Data Classification", 2012, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 10

[9] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, 2017, pp- 215-227 .

[10] Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

[11] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", vol. 8, 2014, pp. 142-147.

[12] Chew Li Sa, Dayang Hanani bt. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossin (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1-6.

[13] Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi (2013), "Predicting Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013), vol. 23, 2013, pp. 32-38.

[14] K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, 2015, pp. 1023-1028.

[15] Bala Sundar V, T Devi and N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications, vol. 48, 2012, pp. 423-428.

[16] Daljit Kaur and Kiran Jyot (2013), "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, vol. 2 2013, pp. 724-729.