Analysis of Prediction Techniques

Rahul Kumar Yadav
Research Scholar
yrahuly@gmail.com
Raja Balwant Singh
Technical Campus
Bichpuri, Agra, UP

Ashok Kumar
Deptt. Of CSE,
R B S Engg
Technical Campus,
Bichpuri, Agra,

Brajesh Kumar Singh
Deptt. Of CSE,
Raja Balwant Singh Engineering
Technical Campus
Bichpuri, Agra, UP

ABSTRACT

The technique through which important information is extracted from the raw data collected within the databases is known as data mining. The future scenarios related to current data can be predicted with the help of prediction analysis technique provided by data mining. The prediction analysis is the combination of clustering and classification. There are numerous techniques proposed by various researchers in order to provide prediction analysis. In this review paper, various techniques proposed by various authors are analyzed to understand latest trends in the prediction analysis.

KEYWORDS

Classification, Clustering, K-means, SVM.

1. INTRODUCTION

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. In data mining there are various data mining tools available which are used to analyze different types of data. Decision making, market basket analysis, production control, customer retention, scientific discovers and education systems are some of the applications that use data mining in order to analyze the collected information [1]. The clustering is the approach which is applied to cluster similar and dissimilar type of data. The clusters are generated by analyzing similar patterns of the input data. 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. The unsupervised data clustering classification method creates clusters and groups 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 cluster [2]. Objects within a class have high resemblance to each other and the objects which are divided into separate classes are more distinct. Following are some broader categories into which the clustering methods have been categorized:

a. 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].

b. 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.

c. 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, these methods can only be helpful while identifying the spherical shaped clusters. It is difficult to obtain arbitrary shaped using the technique of density based clustering.

d. 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.

1.1. Classification in Data Mining

The group membership for data instances can be predicted with the help classification technique within the data mining [5]. Prediction analysis is the process in which outcome will be predicted on the basis of current data. For example, on the basis of current weather information it will be analyzed that day can be either "sunny", "rainy "or "cloudy.

Two steps are followed within this process. They are:

a. 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.

b. Model usage: The second step used in the classification is model usage. In order to classify the test data, the training set is designed of the unknown from the unknown data for the accuracy analysis [6]. 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.

1.2 SVM classifier

For regression, classification as well as general pattern recognition, the SVM classifier is proposed. Due to its high generalization performance without requiring any priori knowledge to add in it, this classifier is considered to be good in comparison to other classifiers. 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 identification of the best classification function. The best classification function metric can be represented geometrically as well [7]. The hyperplane $f(x)$ is separated through the linear classification function for the linearly separable dataset. This hyperplane passes through the middle of two classes which can be said to separating them. The new data instance x_n is classified by testing the sign function 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.

It is the main objective of SVM 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. The amount of space or distance amongst two classes is known as hyperplane. The shortest distance between the closest data points to a point on the hyperplane is known as margin. This can further help us in defining the way to extend the margin which can help in selecting only a few hyperplanes for the solution to SVM even when so many hyperplanes are available [8].

The objective of SVM is to produce linear function which can help in identifying the target function. This can further help in extending the SVM for performing regression analysis. The error models are of quiet help here for the SVRs. 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, the support vectors can be studied. The insensitivity to the outliers can be of beneficial 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 for 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 a set of instances in the program.

This review paper is based on the prediction analysis which is generally done with the classification techniques.

This paper is organized such that in the section 1, the introduction of the prediction analysis is given with various classification techniques. In the section 2, the literature survey is written on the prediction analysis. In the section 3, the result evaluation is described in which number of papers published in IEEE or Springer is studied.

2. Literature Review

Min Chen, et.al [9] proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. The data was gathered from a hospital which included within it, 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.

Akhilesh Kumar Yadav, et.al presented an analysis of different analytic tools that have been used to extract information from large datasets such as in medical field where a huge amount of data is available [10]. The proposed algorithm has been

tested by performing different experiments on it that gives excellent result on real data sets. In real world problem enhanced results are achieved using proposed algorithm as compared to existing simple k-means clustering algorithm.

Sanjay Chakrabotry et.al, (2014) presented clustering tool analysis for the forecasting analysis [11]. The weather forecasting has been performed using proposed incremental K-mean clustering generic methodology. The weather events forecasting and prediction becomes easy using modeled computations. Towards the end section, the authors have performed different experiments to check the proposed approach's correctness.

Chew Li S. et.al, (2013) presented [12] that the results of a particular university's students have been recorded to keep a track using Student Performance Analysis System (SPAS). The design and analysis has been performed to predict student's performance using proposed project on their results data. The data mining technique generated rules that are used by proposed system provide enhanced results in predicting student's performance. The student's grades are used to classify existing students using classification by data mining technique.

Qasem A. et.al, (2013) suggested that the data analysis prediction [13] is considered as important subject for forecasting stock return. The future data analysis can be predicted through past investigation. The past historical knowledge of experiments has been used by stock market investors to predict better timing to buy or sell stocks. There are different available data mining techniques amongst which, a decision tree classifier has been used by authors in this work.

K.Rajalakshmi et.al, (2015) presented study related to [14] medical fast growing field authors. In this field every single day, a large amount of data has been generated and to handle this much of large amount of data is not an easy task. By the medical line prediction based systems, optimum results are produced using medical data mining. The K-means algorithm has been used to analyze different existing diseases. The cost effectiveness and human effects have been reduced using proposed prediction system based data mining.

BalaSundar V et.al, (2012) examined [15] real and artificial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique in order 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 first step is random initialization of whole data, and then a cluster k is assigned to each cluster. The proposed scheme of integration of clustering has been tested and its results show that the highest robustness, and accuracy rate can be achieved using it.

Daljit Kaur et.al (2013) explained [16] that data that contains similar objects has been divided using clustering. The data that contains similar objects is clustered in same group and the dissimilar objects are placed in different clusters. The proposed algorithm has been tested and results show that this algorithm is able to reduce efforts of numerical calculation and complexity along with maintaining an easiness of its implementation. The proposed algorithm is also able to solve dead unit problem.

Authors	Techniques / Algorithms	Datasets	Attributes	Tools Used	Shortcoming	Results
Min Chen, et.al	Naïve Bayesian, KNN and Decision tree	Heart Diseases	79	MATLAB	This classifier has high complexity.	Decision tree performs better in comparison to other classifiers.
Akhilesh Kumar Yadav, et.al	Foggy K-mean Algorithm	Lung cancer Data	9	WEKA	Complexity is high.	Foggy k-mean performs well as compared to K-means
Sanjay Chakrabotry et.al	Incremental k-mean clustering Algorithm	Air pollution Data	7	WEKA	Accuracy is less	The accuracy of proposed method is achieved up to 83.3 percent.
Chew Li S. et.al	BF Tree classifier	Student's Performance	9	WEKA	Complexity is high which increases the execution time.	BF Tree performs well as compared to other tree classifiers
Qasem A. et.al	Decision tree	STOCK Data Prediction	170	WEKA	Accuracy is less which can be increased.	C4.5 classifier performs well as compared to ID3

Table 1: Comparison of Various Techniques

Conclusion

Prediction analysis is the technique of data mining which is used to predict future from the current data. Prediction analysis is the combination of clustering and classification. Clustering algorithm groups the data according to their similarity and classification algorithm assigns 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 various techniques of prediction analysis from where problem is formulated. The formulated problem can be solved in future to increase accuracy of prediction analysis.

References

[1] AbdelghaniBellaachia and ErhanGuven (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] AzharRauf, Mahfooz, Shah Khusro and HumaJaved (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] AzharRauf, Sheeba, SaeedMahfooz, Shah Khusro and HumaJaved (2012), "*Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity*," Middle-East Journal of ScientificResearch, 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, volume 3, issue 12, pp- 551-559, 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, YixueHao, 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, DivyaTomar and SonaliAgarwal (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., Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (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 EmanAlnagi (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] BalaSundar V, T Devi and N Saravan, (2012) "*Development of a Data Clustering Algorithm for Predicting Heart*", International Journal of Computer Applications, vol. 48, 2012, pp. 423-428.

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