Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 5, May 2019, pg.15 – 22

Prediction Analysis Techniques of Data Mining: A Review

Himani Rani
Research Scholar, Punjabi University, Patiala, duahimani13@gmail.com

Dr. Gaurav Gupta
Assistant Professor, Punjabi University, Patiala, gaurav.shakti@gmail.com

ABSTRACT: Data mining within the databases is called a technique from which the extraction of necessary information can be done from the raw information. With the help of the prediction analysis technique provided by the data mining the future scenarios regarding to the current information can be predicted. The prediction analysis is the combination of clustering and classification. In order to provide prediction analysis there are several techniques presented through many researchers. 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

Data mining is the patterns for analyzing information and the process to extract the interesting knowledge. In data mining, various data mining tools available which are used to analyze different types of data. For analyzing the data information few applications which is used by data mining are such as making decisions, analysis on market basket, production control, and customer retention, scientific discovers and education systems [1]. Applied to similar cluster and not same type of data is referred to clustering in this approach.The clusters are generated by analyzing similar patterns of the input data. While categorizing genes with same functionality and in population gain insight into structures can be inherited in biology for deriving plant and animal taxonomies. In 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 objects as these in different clusters are distinct and that are in same cluster are very similar to each other. In data mining, cluster analysis is considered a traditional topic which is applied for the knowledge discovery. The data objects are grouped as a set of disjoint classes which are known as cluster

[2]. Objects which are divided into separate classes are more different and within a class objects have high resemblance to each other. In order to determine patterns and predicting future outcomes and trends predictive analytics is the practice of extracting from existing data sets. Future predictions are not provided through prediction analysis. In the future with an acceptable level of reliability includes what-if scenarios and risk assessment forecast is provided by the prediction analysis. Future possibilities are completely predicted through the prediction analysis. In order to better understand customers, products and partners and to identify potential risks and opportunities for company predictive models are used to analyze current data and historical facts applied to business. To make future business forecasts including data mining, statistical modeling and machine learning to help analysts it uses a number of techniques. From data Predictive analytics is an area of statistics that deals with extracting information and used it for predicting trends and behavior patterns. Calculation of statistical probabilities of future events online is the enhancements of predictive web analytics. Data modeling, machine learning, AI, deep learning algorithms and data mining are included in the Predictive analytics statistical techniques predictive analytics can be applied to any type of unknown whether it be in the past, present or future often the unknown event of interest is in the future. To predict the likely behavior of individuals, machinery or other entities Predictive analytics software applications use variables that can be measured and analyzed. Such as age, gender, location, type of vehicle and driving record, pricing and issuing auto insurance policies are taken in an account potential driving safety variables through the insurance company. With statistical methods and the ability to build predictive data models Predictive analytics requires a high level of expertise. it's typically the domain of data scientists, statisticians and other skilled data analysts which is the complete outcomes of prediction analysis. For helping to gather relevant data and prepare it for analysis is supported through data engineers. Therefore, with data visualization, dashboards and reports are supported through software developers and business analysts. Clustering methods divided into categories are as follows:

a. **Partitioning Methods: -** the basic functioning of this method is the collection of the samples in a way to generate clusters of same objects that are of high similarities. 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 classification of this method is done with the involvement decomposition. It is divisive and agglomerative methods based upon [4]. Agglomerative technique is the bottom up technique at which the first step is the formation of the separate group. Merging is done when the groups are near to each other.

c. **Density Based Methods: -** In many techniques the distance amongst the objects is taken for the separation of the objects into clusters as a base into clusters. 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: -** It is known as the generation of grid structure by the quantizing the space of the object to the finite number of cells. This method is independent as it is not dependent on the availability of the number of data objects and also has a high speed.

1.1. Classification in Data Mining

Within the data mining the prediction of the group membership for instance information can be done with the help of the classification technique [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 explains the group of classes of predetermined. Wide numbers of tuples are utilized in the construction of the model known as training set. Classification of the rules, decision trees or mathematical formulae/regression is shown in this method.

b. Model usage: The second way 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 result of the classification of the model is used to compare in sample test with a label that is known. Test set is not dependent on training set.

1.2 SVM classifier

In this study the author proposed SVM classifier for regression, classification and also the general pattern recognition. Due to its high generalization performance without requiring any prior knowledge to add in it, this classifier is considered to be good in comparison to other classifiers. The performance is even better such as extremely high of the input space dimension. The SVM requires best classification function identification for differentiating of training data between the two classes The classification function metric may represent in a geometric manner as well [7]. The hyper plane *f(x)* is separated through the linear classification function for the linearly separable dataset. This hyper plane passes through the middle of two classes which can be said to separating them. x_n is classified by testing the sign function of the new data instance function $f(x_n)$; x_n which refers to the positive class if $f(x_n) > 0$. This is done after the determination of a new function.

Determination of the best function by increasing the margin between the two classes is an important objective of SVM. There are many linear hyper planes because of this fact. Hyper plane is amongst the two classes an amount of space or distance present. Margin is closest between the closest data points to a point with a shortest distance on the hyper plane**.** This can further help us in defining the way to extend the margin which can help in selecting only a few hyper planes for the solution to SVM even when so many hyper planes are available [8].

For an identification of the target function the aim of the SVM is to produce linear function. Performance of the regression analysis can help to extend the SVM. The error models are of quiet help here for the SVRs. Within an epsilon amount the error is defined zero of the differences between real and predicted values.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] presented on the basis of multimodal disease risk prediction (CNN-MDRP) algorithm called a novel convolution neural network**.** 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 comparison with existing simple k-means clustering algorithm using the algorithm results are achieved in real world problem.

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.

Ming, J, et.al (2018) proposed multi-dimensionality and nonlinearity the Characteristics of the technical and economic data of mining enterprises. Using technologies of big data analysis and data mining the analysis method of the technical and economic data is researched. Simplification of the fluctuation pattern and influencing factors of the mineral products price are done. Using artificial neural network the prediction model of the mineral products price is established [17]. The prediction model of the geological missing data is established on the basis of techniques of geo statistics and artificial neural network. Regularity of geological data of group boreholes and of geological data of all boreholes the regularity is discussed and analyzed by using the model. The practicability of the prediction model is strong, and the prediction accuracy is high as shown in the outcomes of the proposed approach through the authors. Due to the limitation of technical conditions and equipment conditions during the process of mineral development there is a loss of a lot of geological data that decreases accuracy of the ore body shape and that of reserves estimation in this study.

Sakhare, A. V, et.al (2017) proposed in data mining paper shows a survey of road accident analysis methods an important role played in transportation is the system road accident analysis. Using the different methods of data mining this paper Road Accident Data Analysis is described. The study of K-mean algorithm is given in this paper. Clusters are created and analyze them with the help of SOM [18]. It is used as an unsupervised learning method based on neural network known as self organizing method. Analysis accuracy is improved through this. Because no. of people death and injured for that improve the road transportation system is needed in our daily life there are no. of accident increases and it is big problem to us. For

finding a no. of pattern to analysis the road accident data which help to find prediction of accident reasons and improve the accuracy of analysis compare to k-means clustering algorithm is known as the research self organization map (SOM).

Chauhan, C, et.al (2017) proposed to analyze the victim system where the attack is occurring and also the forensic kit tool generates the file and analyzes the data in this proposed approach. This approach can analyze previously unknown, useful information from an unstructured data using the concept of data mining [19]. For the identification of criminal and it has been found to be pretty much effective in doing the same Predictive policing means, using analytical and predictive techniques. Methodical approach for identifying and analyzing patterns and trends in crime is the Crime analysis. Crime data analysts can help the Law enforcement officers to speed up the process of solving crimes with the increasing origin of computerized systems. During analysis of experimental data it is concluded that advanced ID3 algorithm is more reasonable and more effective classification rules

Anoopkumar M, et.al (2016) proposed give a comprehensive survey towards the research papers which would have discussed different Data Mining Methods especially the mostly utilized and trendy algorithms applied to EDM context. For computing educators and professional bodies this paper accumulates and relegates literature, identifies consequential work and mediates. In this field to date this paper conducted a comprehensive study on the recent and relevant studies kept through [20]. Developing models for improving academic performances and improving institutional effectiveness is the main focus of this study on methods of analyzing educational information. An interdisciplinary ingenuous research area that handles the development of methods for exploring information arising in scholastic fields is known as the Educational Data Mining (EDM). For effective education planning as a result, it provides intrinsic knowledge of teaching and learning process. Ameliorating pedagogical process, presaging student performance, comparison of the precision of data mining algorithms, and demonstrate the maturity of open source implements are the outcomes of these studies give insight into techniques in this proposed approach.

Lee, E., Jang, et.al (2018) proposed using commercial game log data competition framework for game data mining in this paper. Promoting the research of game data mining by providing commercial game logs to the public is the purpose of the game data mining competition. From other types of game AI competitions that targeted strong or human-like AI players and content generators the goal of the competition was very different [21]. With external researchers game companies avoid sharing their game data this approach enabled researchers to develop and apply state-of-the-art data mining techniques to game log data. To predict whether a player would churn and when the player would churn during two periods between which the business model was changed to a free-to-play model from a monthly subscription was the main objective of this proposed approach. Highly ranked competitors used deep learning; tree boosting and linear regression was the outcome of the competition revealed in this proposed approach by the researchers and authors.

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
K.Rajalakshmi	Medical fast growing field	Prediction based systems	3	Python	A large amount of data has been generated and to handle this much of large amount of data	The cost effectiveness and human effects have been reduced using proposed prediction system based data mining.
BalaSundar	real and artificial datasets	to predict diagnosis of heart diseases	5	WEKA	The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis	Show that the highest robustness, and accuracy rate can be achieved using it.
Daljit Kaur	contains similar objects has been divided using clustering	dissimilar objects	12	Python	algorithm is able to reduce efforts of numerical calculation and complexity	The proposed algorithm is also able to solve dead unit problem.
Ming, J	multi-dimensionality and nonlinearity the Characteristics of the technical	technical and economic data	2	MATLAB	Simplification of the fluctuation pattern and influencing factors of the mineral products price are done	during the process of mineral development there is a loss of a lot of geological data that decreases
Sakhare	a survey of road accident analysis methods an important role played in transportation	Road Accident Data Analysis	2	WEKA	Clusters are created and analyze them with the help of SOM	improve the accuracy of analysis compare to k-means clustering algorithm
Anoop kumar	different Data Mining Methods especially the mostly utilized	comprehen sive survey	5	MATLAB	In this field to date this paper conducted a comprehensive study on the recent and relevant studies	comparison of the precision of data mining algorithms, and demonstrate the maturity of open source implements are the outcomes of these studies
Lee, E.,	Commercial game log data competition framework was used for game data mining	d tested on the game log data of Blade & Soul of NCSOFT	3	MATLAB	To predict whether a player would churn and when the player would churn during two periods between which the business model was changed to a free-to-play model from a monthly subscription was the main objective of this proposed approach.	Highly ranked competitors used deep learning; tree boosting and linear regression was the outcome of the competition revealed in this proposed approach by the researchers and authors.

Table 1: Comparison of Various Techniques

Conclusion

Future prediction is done from the current information by the prediction analysis which is the technique of data mining. The combining of clustering and classification is known as the prediction analysis. Clustering algorithm groups the data according to their similarity and classification algorithm assigns class to the data. In terms of many parameters several prediction analysis algorithms are reviewed and analyzed in this paper. 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 "Predicting Stock Prices Using Data Mining Techniques", the International Arab Conference on Information Technology (ACIT'2013), vol. 23, 2013, pp. 32-38, (2013),

[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

[17] Ming, J., Zhang, L., Sun, J.& Zhang, Y, "Analysis models of technical and economic data of mining enterprises based on big data analysis", International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, IEEE, 3rd

[18] Sakhare, A. V., & Kasbe, P. S "A review on road accident data analysis using data mining techniques", International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017
[19] Chauhan, C., & Sehgal, S, "A review: Crime analysis using data mining techniques and algorithms", International Conference on Computing, Communication and Automation (ICCCA), 2017
[20] Anoopkumar M, & Rahman, A. M. J. M. Z, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration, International Conference on Data Mining and Advanced Computing (SAPIENCE), (2016)
[21] Lee, E., Jang, Y., Yoon, D.-M., Jeon, J., Yang, S., Lee, S, "Kim, K.-JGame Data Mining Competition on Churn Prediction and Survival Analysis" using Commercial Game Log Data Transactions on Games, IEEE, 2018

