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. 4, April 2019, pg.187 – 191

Prediction Analysis Techniques of Data Mining: A Review

Kaveri Giri
Research Scholar, Ganga Institute of Technology and Management, Jhajjar, Haryana
kav113giri@gmail.com

Dheer Dhwaj Barak
Associate Professor, Ganga Institute of Technology and Management, Jhajjar, Haryana
barakdheer410@gmail.com

ABSTRACT: The data mining is the technology which can mine the useful information from the rough data. The prediction analysis is the technique of data mining which can predict the future situations from the current data. The prediction analysis is the combination of the clustering and classification. In this review paper, various techniques which are used for the prediction analysis are reviewed and analyzed in terms of various parameters.
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 customer categorized group and purchasing patterns done by clustering can be used by marketer to discover their customer's interest. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. 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 that classify all documents available on Web. 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 topics that is the first step in discovery of knowledge. The data objects are grouped into a set of disjoint classes which is known as clusters

[2]. Now, objects within a class have high resemblance to each other and in the meantime objects in separate classes are more unlike. 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 technique. However, these methods can only be helpful while identifying the spherical shaped clusters. It is difficult to obtain arbitrary shaped clusters within these techniques.

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. In order to predict the data for example classification can be utilized by the applications on a specific day to identify the weather which can be either "sunny", "rainy" or "cloudy". Two steps are followed within this process. They are [5]:

a. Model Construction: Model construction describes the set of predetermined classes. The class label attribute determined 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.

b. Model usage: second step used in the classification is model usage. In order to classify the future and unknown objects model usage is widely used as this model estimates the accuracy of the model. 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

SVM stands for support vector machine. It is a binary classifier that maximizes the margin. The best hyperplane which separates all the data points of an individual class can be identified through the classification provided by SVM. The largest margin between the two classes describes the best hyperplane for an SVM [6]. The maximum width between the slabs parallel to the hyperplane is known as margin which has no interior data points. The svm algorithm is used to separate maximum margin in hyperplane. The margin planes determined using the point from each class are called support vectors (SVs). It has many applications such as bioinformatics, text, image recognition etc. It becomes popular due to its success in handwritten digit recognition. A huge and varied community works on them like machine learning, optimization, statistics, neural networks, functional analysis, etc.

2. Literature Review

Min Chen, et.al proposed in this paper [7], 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 types of data. In order to make predictions related to the chronic disease that had been spread within 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 in this paper [8], that different analytic tool has been used to extract information from large datasets such as in medical field where a huge data is available. 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) stated in this paper [9], that powerful tool clustering is used as different forecasting tools. 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. In the last the authors have performed different experiments to check the proposed approach correctness.

Chew Li S. et.al, (2013) presented in this paper [10] particular university student results has 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 to gives enhanced results in predicting student performance. The student's grades are used to classy existing student using classification by data mining technique.

Qasem A. et.al, (2013) presented in this paper [11] that data analysis prediction is considered as import subject for forecasting stock return. The data analysis future 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 out of all a decision tree classifier has been used by authors in this work.

K.Rajalakshmi et.al, (2015) presented in this paper [12] a study related to 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. The medical line prediction based systems optimum results are produced by medical data mining. The K-means algorithm has been used to analyze different existing diseases. The cost effectiveness and human effects has been reduced using proposed prediction system based data mining.

Bala Sundar V et.al, (2012) examined in this paper [13] real and artificial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique results 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 then a cluster k is assigned to each cluster. The proposed scheme of integration of clustering has been tested and its results show that highest robustness, accuracy rate can be achieved using it.

Daljit Kaur et.al (2013) explained in this paper explained [14] that data contained similar objects has been divided using clustering. A data of similar objects are in same group and in case dissimilar objects occur then it will be compared with other group's objects.. The proposed algorithm has been tested and results shows that it 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.

Author	Year	Description	Outcomes
Min Chen, et.al	2017	A novel convolution 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 types of data. In order to make predictions related to the chronic disease that had been spread within 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	2013	In this paper, that different analytic tool has been used to extract information from large datasets such as in medical field where a huge data is available. The SGPGI real data set has been used that are always linked with different challenges. The classification becomes inefficient due to noise, high dimensional and missing values.	The proposed algorithm has been tested by performing different experiments on it that gives excellent result on real data sets.
Sanjay Chakrabotry et.al	2014	In this paper author suggest that powerful tool clustering is used as different forecasting tools. The weather forecasting has been performed using proposed incremental K-mean clustering generic methodology. The purpose behind this paper is to analyze air pollution for it they have used dataset of west Bengal. The clusters peak mean values are used to develop a weather category list and K-means clustering is applied on the dataset of air pollution.	The weather events forecasting and prediction becomes easy using modeled computations. In the last the authors have performed different experiments to check the proposed approach correctness.
Chew Li S. et.al	2013	In this paper particular university student results has 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 student's grades are used to classy existing student using classification by data mining technique.
Qasem A. et.al	2013	In this paper, data analysis prediction is considered as import subject for forecasting stock return. The data analysis future 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 out of all a decision tree classifier has been used by authors in this work.

Table 1: Table of Comparison

3. Problem Formulation

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. The CNN-MDRP modal for the prediction analysis is based on the neural networks. In which the clustering algorithm called k-means clustering is applied which can categorize the data into certain number of classes. The clustered data is given as input to the classification algorithm which can divide the dataset into two parts testing and training. The SVM classifier is used to classify the data into certain number of classes. In the k-mean clustering algorithm, the centered points are calculated by taking arithmetic mean of the whole dataset which can reduce accuracy of prediction analysis. When the dataset is complex, it is difficult to establish relationship between the attributes of the dataset. In this research

work, improvement in the k-mean clustering will be applied which can select centered points in the efficient manner to categorize input data. The proposed improvement directly increase accuracy of clustering and reduce execution time.

Conclusion

In this paper, it is concluded that 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 group 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 from where problem is formulated. The formulated problem can be solved in future to increase accuracy of prediction analysis.

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] Kajal C. Agrawal and Meghana Nagori (2013), "*Clusters of Ayurvedic Medicines Using Improved K-means Algorithm*", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, 2013, pp. 546-552.
[7] 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
[8] 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.
[9] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "*Weather Forecasting using Incremental K-means Clustering*", vol. 8, 2014, pp. 142-147.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.

