Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Prediction Analysis in Data Mining: A Review

Shalu Saxena
Department of Computer Science & Engineering,
PG Scholar, SRMCEM
Lucknow

Dr. Pankaj Kumar
Department of Computer Science & Engineering,
Assistant Professor, SRMCEM
Lucknow

Dr. Raj Gaurang Tewari
Department of Computer Science & Engineering,
Assistant Professor, SRMCEM
Lucknow

***Abstract*:- The data mining is the technology which is applied to extract the useful information from the rouge information. The clustering is the efficient technique which is applied to cluster the similar and dissimilar type of data. The prediction analysis is applied in which technique of clustering is applied which will cluster the data and in the second step technique of classification is applied which will classify the similar and dissimilar type of data. In this paper, various technique of prediction analysis is reviewed and discussed in terms of various parameters**

INTRODUCTION

1.1 DATA MINING

Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge. Data mining is used for analysis purpose to analyze different type of data by using available data mining tools. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [1].

1.2 CLUSTERING

Cluster analysis [11] has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. Data clustering [14] (or just clustering), is an unsupervised classification method aims at creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. Clustering is the procedure of grouping data objects into a set of disjoint classes, called clusters.

1.3 K-MEAN CLUSTERING

The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, (*C1 …… Ck*), such that the sum of the squared distances of each data point, *xi* ,$1 \leq i \leq n$, *to* its nearest cluster centre *Cj*, $1 \leq j \leq k$, is minimized.

1.4 CLASSIFICATION

Classification is a data mining technique which comes under machine learning technique to predict group membership for data instances. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity. There are number of classifiers available for classification techniques.

A. SVM Classifier
B. Decision Tree Induction
C. Bayesian Networks
D. K-Nearest Neighbours
E. Instance Based Learning

A. SVM Classifier: Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

B. Decision Tree Induction: Decision Tree Induction by sorting them based on the feature values. Each node in a decision tree represents a feature in an instance to be classified and each branch represents a value that a node can assume. Instances are sorted on the basis of the feature values. The feature that best divides the training data would be the root node of the tree. There are numerous methods for finding the feature that best divides the training data such as information gain.

Greedy is a basic algorithm for a decision tree induction that constructs decision tree according to divide and conquer scheme in a top-down manners. People can easily understand decision tree classifiers are belongs to a specific class. Since a decision tree constitutes a hierarchy of tests, an unknown feature value during classification is usually dealt with by passing the example down all branches of the node where the unknown feature value was detected, and each branch outputs a class distribution. The output is a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

combination of the different class distributions that sum to 1. The assumption made in the decision trees is that instances belonging to different classes have different values in at least one of their features. Decision trees tend to perform better when dealing with discrete/categorical features.

C. Bayesian Networks: A Bayesian Method is a graphical model to find out relationship among variable features. The Bayesian Network structure S is a directed acyclic graph and nodes are in one-to-one correspondence relationship among their features. In addition to it, a feature node is conditionally independent from its non-descendents given its parents. The task of learning a Bayesian networks can be divided into two subtasks. First is learning of the DAG structure of the network and other is determination of its parameters.

Probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents. Given the independences encoded into the network, the joint distribution can be reconstructed by simply multiplying these tables. Within the general framework of inducing Bayesian networks, there are two scenarios: known structure and unknown structure.

The most interesting feature of BNs, compared to decision trees or neural networks, is most certainly the possibility of taking into account prior information about a given problem, in terms of structural relationships among its features. This prior expertise, or domain knowledge, about the structure of a Bayesian network can take the following forms: 1. Declaring that a node is a root node, i.e., it has no parents. 2. Declaring that a node is a leaf node, i.e., it has no children. 3. Declaring that a node is a direct cause or direct effect of another node. 4. Declaring that a node is not directly connected to another node. 5. Declaring that two nodes are independent, given a condition-set. 6. Providing partial nodes ordering, that is, declare that a node appears earlier than another node in the ordering. 7. Providing a complete node ordering. A problem of BN classifiers is that they are not suitable for datasets with many features. The reason for this is that trying to construct a very large network is simply not feasible in terms of time and space. A final problem is that before the induction, the numerical features need to be discredited in most cases.

D. K-Nearest Neighbor Classifier: Nearest Neighbor classifier are based on by analogy. The n dimensional numeric attributes are described by the training samples. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample.

The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space.

Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. This contrasts with eager learning methods, such a decision tree induction and back-propagation, which construct a generalization model before receiving new samples to classify. Lazy learners can incur expensive computational costs when the number of potential neighbours (i.e. stored training samples) with which to compare a given unlabeled sample is great. Therefore, they require efficient indexing techniques. Expected lazy learning methods are faster than training than eager methods, but slower at classification since all computation is delayed to that time. Unlike decision tree induction and back propagation, nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data.

Nearest neighbor classifiers can also be used for prediction, that is, to return a real-valued prediction for a given unknown sample. In this case, the classifier returns the average value of the real-valued associated with the k nearest neighbors of the unknown sample. The k-nearest neighbors' algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidian distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

E. Instance Based Learning: Another and last category is instance based learning. This category comes under statistical methods. Lazy-learning algorithms require less computation time during the training phase than eager-learning algorithms but more computation time during classification process. One of the most straightforward instance-based learning algorithms is the nearest neighbor algorithm. K-Nearest Neighbor (KNN) is based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbors. The KNN locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label.s

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

2 LITERATURE REVIEW

Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", 2013 : In this paper they explained that huge data is available in medical field to extract information from large data sets using analytic tool. In this paper a real data set has been taken from SGPGI. Real time data sets are always interlinked with some challenges like missing values, high dimensional values and noise etc which is not efficient for all the classification. Therefore clustering is the alternate solution for data analytics. The main focus of this paper is to develop a novel technique based upon foggy k-mean clustering. The result of the experiment depicts that foggy k-means clustering algorithm has excellent result on datasets which are real as compared to simple k-means clustering algorithm and provides a enhanced result to the real world problem.

Sanjay Chakrabotry et.al, "Weather Forecasting using Incremental K-means Clustering", 2014: In this paper they explained that clustering is the powerful tool which used in various forecasting tools. In this paper generic methodology of incremental K-mean clustering is proposed for weather forecasting. This research has been done on air pollution of west Bengal dataset. This paper generally uses typical K-means clustering on the main air pollution database and a list of weather category will be developed based on the peak mean values of the clusters. Whenever new data are coming, the incremental K-means is used to group data into those clusters where weather category has been already defined. Thus it is able to predict weather information of future. This forecasting database is totally based on the weather of west Bengal and this forecasting methodology is prepared to mitigate the consequences of air pollutions and launch focused modeling computations for prediction and forecasts of weather events. Here correctness of this approach is also measured.

Chew Li Sa et.al, "Student Performance Analysis System", 2013: In this paper they proposed a system named Student Performance Analysis System (SPAS) to keep track of student's result in a particular university. The proposed project offers a system which predicts performance of the students on the basis of their result on the basis of analysis and design. The proposed system offers student performance prediction through the rules generated via data mining technique. The data mining technique used in this project is classification, which classifies the students based on students' grade.

Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", 2010: In this paper they presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project. This preprocessed dataset consists of 151, 886 records which have 16 fields from the SEER. After that they have investigated three data mining techniques: Naïve Bayes, back propagated neural network and C4.5 decision tree algorithm. Several experiments have been implemented using above mentioned experimental. At the end existing techniques have been compared with the achieved prediction performance. Later on it is concluded that C4.5 algorithm has a much better performance than other two techniques.

Qasem A. et.al, "Predicting Stock Prices using data mining techniques", 2013: In this paper they explained that forecasting stock return is one the important subject to be learn for prediction for data analysis. It is analysis that past investigations help to predict future in data analysis. In this paper they try to help investors in stock market better timing for the buying and selling stocks on the basis of knowledge of past historical experiments. In this paper they define decision tree classifier which is one of the best data mining techniques.

K.Rajalakshmi et.al, "Comparative Analysis of K-Means Algorithm in Disease Prediction", 2015

In this paper they represented an extremely fast growing field of medical. A huge amount data has been generated by this field every day. To handle this data is very difficult, so there is a need of a technology to handle this data. To turn these data into useful pattern, there is a need of a data to be mined. The medical data mining are useful to produce optimum results on prediction based system of medical line. This paper analyzes various disease predictions techniques using K-means algorithm. This data mining based prediction system are reduces the human effects and cost effective one.

Oyelade et.al, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", 2010: In this paper they defined the ability of the student performance of high learning. To analyze student result based on cluster analysis and use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was combined with deterministic model to analyze student's performance of the system.

Bala Sundar V et.al, "Development of a Data Clustering Algorithm for Predicting Heart", 2012:
In this paper they examined the conclusion of the accuracy of the result by using k-mean clustering technique in prediction of heart disease diagnosis with real and artificial datasets. Clustering is the method of cluster analysis which aims to cluster to partition into k clusters and each cluster has its observations with nearest mean. Each cluster assigned to the cluster k and started from random initialization. The proposed technique further divided into k groups. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid. The research result shows that the integration of clustering gives promising results with highest accuracy rate and robustness.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

AUTHOR	YEAR	DESCRIPTION	OUTCOMES
Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal	2013	A real data set has been taken from SGPGI which are always interlinked with some challenges like missing values, high dimensional values and noise etc. So foggy K-mean clustering is developed to create accurate cluster.	Foggy K-mean clustering algorithm gives better result on real dataset . It provides better solution to real world problem.
Sanjay Chakrabotry et.al	2014	Incremental K-mean clustering is proposed for weather forecasting. Weather category will be based on peak mean values of the clusters.	Mitigate the consequences of air pollution and launch focused modelling computation for prediction and forecast of weather events.
Chew Li Sa et.al	2013	They proposed a system named Student performance analysis system(SPAS) to keep track of students result in a particular university.	The proposed system offers student performance prediction via data mining technique of classification which classifies the student based on student's grade.
Qasem A. et.al	2013	In this paper they explained that forecasting stock return is one the important subject to be learn for prediction for data analysis.	They try to help investors in stock market better timing for the buying and selling stocks on the basis of knowledge of past historical experiments. They define decision tree classifier which is one of the best data mining techniques
K.Rajalakshmi et.al	2015	An extremely fast growing field of medical. A huge amount of data has been generated by this field every day. To turn these data into useful pattern there is a need of a data to be mined.	This paper analyses various disease prediction techniques using K-mean algorithm. This technique reduces the human effects and cost effective one.
Bala Sundar V et.al	2012	In this paper they examined the conclusion of the accuracy of the result by using k-mean clustering technique in prediction of heart disease diagnosis with real and artificial datasets.	The research result shows that the integration of clustering gives promising results with highest accuracy rate and robustness.

CONCLUSION

In this paper, it is been concluded that prediction analysis is the technique which is applied to predict the future scope. The prediction analysis contains the two steps in the first step technique of clustering is applied which will cluster similar type of data. In the second step technique of classification is applied which will classify the data. In this paper, various technique of prediction analysis is reviewed and discussed in future prediction analysis techniques will be improved for the performance improvement

REFRENCES

[1] Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT) 2013

[2] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-means Clustering", 2014

[3] Chew Li Sa; Bt Abang Ibrahim, D.H.; Dahliana Hossain, E.; bin Hossin, M., "Student performance analysis system (SPAS)," in *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on* , vol., no., pp.1-6, 17-18 Nov. 2014

[4] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, 2010

[5] QASEM A. AL-RADAIDEH, ADEL ABU ASSAF 3EMAN ALNAGI, " Predictiong Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[6] K.Rajalakshmi, Dr.S.S.Dhenakaran,N.Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015

[7] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010

[8] Bala Sundar V,T Devi, N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012

[9] Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013

[10] Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999.

[11] Azhar Rauf ,Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", *Middle-East Journal of Scientific Research* 12 (7): 959-963, 2012 ISSN 1990-92332012

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

[12] Madhu Yedla, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center*", International Journal of Computer Science and Information Technologies*, Vol. 1 (2) 2010, page 121-125

[13] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, *Proceedings of the World Congress on Engineering* , Vol IWCE 2009, July 1 - 3, 2009, London, U.K

[14] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" *PLoS ONE*, Volume 7, Issue 12, pp-56-62, 2012.

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

[16] Kajal C. Agrawal and Meghana Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," *International Conf. on Advances in Computer Science and Electronics Engineering*, 2013.

[17] K.Ranjini and Dr. N.Rajalingam, "Performance Analysis of Hierarchical Clustering Algorithm ", Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011, 2011

[18] Stan Salvador and Philip Chan, "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms", 2010