IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol. 10, Issue.1, January 2022 E-ISSN: 2347-2693

Overview of the Predictive Data Mining Techniques

C. Ganesh[1*], E. Kesavulu Reddy[2]

[1,2]Dept. of Computer Science, S.V. University College of CM&CS, Tirupati. Andhra Pradesh-India-517502

[*]*Corresponding Author: ekreddysvu2008@gmail.com Tel.: +91 9866430097*

DOI: https://doi.org/10.26438/ijcse/v10i1.2836 | Available online at: www.ijcseonline.org

Received: 05/Jan/2022, Accepted: 25/Jan/2022, Published: 31/Jan/2022

Abstract: Data mining conciliations talented ways to expose secreted designs within huge volumes of data. These hidden designs can possibly be used to prediction forthcoming performance. The descriptive data mining tasks characterize the general properties of the data present in the database, while in contrast predictive data mining technique perform inference from the current data for making prediction. This overview briefly introduces these two most important techniques that perform data mining task as Predictive and Descriptive. Between this predictive and descriptive they consist of their own method as Classification, clustering, Data mining (knowledge discovery from data) may be viewed as the abstraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns and models from observed data or a method used for analytical process designed to explore data. We know Data mining as knowledge discovery. Basically, Extraction or "MINING" means knowledge from large amount of data. 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 paper describes various techniques proposed by various authors are analysed to understand latest trends in the prediction analysis.

Keywords: Extraction, Predictive Techniques, Database, Classification, SVM, Clustering.

1. INTRODUCTION

Data mining is the patterns for analysing information and the process to extract the interesting knowledge. 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. 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.

Data Predictive analytics is an area of statistics that deals with extracting information and used it for predicting trends and behaviour patterns. Calculation of statistical probabilities of future events online is the enhancements of predictive web analytics. Data modelling, 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 behaviour of individuals, machinery or other entities Predictive analytics software applications use variables that can be measured and analysed. 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.

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.

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.

*A. Classification i*n 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. xn is classified by testing the sign function of the new data instance function f(xn); xn which refers to the positive class if f(xn)> 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.

THE SCOPE OF DATA MINING

Data mining originates its name from the resemblances between probing for treasured business information in a large database, for instance, discovering linked products in gigabytes of store scanner data, and pulling out a mountain for a vein of treasured ore. The two processes above will like to find out where exactly the treasured can be found. If the database given is satisfactory in size and quality, then the data mini9ng technology can produce new prospects by given these competences.

Data mining tools swing through databases and categorize formerly hidden designs in one step. An instance of design detection is the investigation of retail sales data to categorize apparently dissimilar products that are often bought together. Other design detection problems include noticing deceitful credit card transactions and recognizing irregular data that could symbolize data entry inputting errors.

Data mining techniques can produce the benefits of mechanization on prevailing software and hardware platforms, and can be applied on new systems as prevailing platforms are elevated and new products established. Data mining tools can investigate large databases in few minutes if they are applied on high performance parallel processing

systems. Processing faster means that users can mechanically trial with additional models to apprehend compound data. High swiftness makes it hands-on for users to investigate massive amounts of data. Databases that are bigger, always in turn produce betterquality forecast.

II. RELATED WORK

He [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.

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.

He explains (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.

He 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 have 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.

He 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.

She 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.

He 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.

He 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.

He 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.

He 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).

He 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.

He 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).

He 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 researcher's 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.

III. DATA MINING MODELS

There are two main data mining models types. These are

1. Predictive
2. Descriptive

The descriptive model recognizes the designs or relationships in data and discovers the properties of the data studied. For instance, Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is like classification however the groups are not predefined, but then again are well-defined by the data alone. It is also referred to as unsubstantiated learning or subdivision. It is the wall off or splitting up of the data into collections or clusters. The clusters are well-defined by learning the performance of the data by the domain experts. The term splitting up is used in very precise framework; it is a process of separation of database into split grouping of related tuples.

Predictive analytics has been defined by [22]as to have data modelling as a prerequisite when making authoritative predictions about the future using business forecasting and simulation. These address the questions of "what will happen?" and "why will it happen?" A different study by [23] defines Predictive analytics as a tool that "uses statistical techniques, machine learning, and data mining to discover facts in order to make predictions about unknown future events," in investigating a domain-specific framework for Predictive analytics in manufacturing. The predictive model makes forecast about unidentified data values by using the identified values. For instance, Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are meant to forecast the forthcoming state of the data.

Forecast is the process of investigating the existing and previous states of the attribute and forecast of its forthcoming state. Classification is a method of plotting the target data to the predefined clusters or classes. The regression includes the book learning of purpose that map data element to actual valued forecast variable. In the time series analysis, the value of an attribute look at it as differs over time. In time series analysis the distance measures are used to define the resemblance between different time series, the structure of the line is studied to define its deeds and the past time series plot is used to forecast forthcoming values of the variable.

A. Descriptive Models

Summarization is the process of giving the recap information from the data. The association rule discovers the connection amongst the diverse traits. Association rule mining is a two-step process: Finding all frequent item sets and generating strong association rules from the frequent item sets. According to Mortenson, Doherty, & Robinson (2014), descriptive analytics recaps and transforms data into expressive information for reporting and one-toone care but also allows for thorough examination to answer

questions such as "what has occurred?" and "what is presently bang upto-date?" SAP, (2014) also described descriptive analytics as control panel applications that support development implementation in sales and procedures administration, allowing for real-time tracking.

Summarization can be observed as squeezing a given set of dealings into a smaller set of designs while recollecting the supreme likely information. Summarization is a common and authoritative though often time-consuming method to examining large datasets. For example, suppose one wants to examine census data in order to appreciate the association amongst level of education and salary in Ghana. A very dense summary of the census can be observed by plotting the average salary by education level. In addition, it may be more seethrough, for instance, to breakdown the average salaries by age group, or to eliminate distant salaries. In general, the summarization involves both identifying overall trends and important exceptions to them, it does not lead to forecast.

There can also be Multidimensional Sequential Pattern Mining as expressed by Pinto, Han, Pei, Wang, Chen, &Dayal, (2001). Let's consider pattern P1= {use a 100-hour free internet access bundle⇒donate to 20 hours'/month bundle⇒elevated to 50 hours per month bundle⇒elevated to unlimited bundle from an Internet Service Provider(ISP) like Vodafone in Ghana. This pattern may hold for all customers below age of 25 who are males. For other consumers, design P2= {use a 100-hour free internet access bundle⇒elevated to 50 hours per month bundle} may hold. Clearly, if successive design mining can be related with consumer group or other multi-dimensional information, it will be more operative since the confidential designs are often more valuable. Pinto et al. (2001) again propose a mixing of well-organized sequential design mining and multi-dimensional investigation procedures (Seq-Dim and DimSeq) and implanting multi-dimensional information into sequences and mine the whole set by means of a uniform sequential design mining technique (Uni-Seq). A multidimensional sequence database has the schema: RID; record identifier, Pi, Pii … Pm; attributes, and S is the sequence. The derived formula only helps to link the items but does not help one to forecast into the future though decision can be taken out of that.

Cluster analysis is another type of Descriptive model which groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. If the similarity (or homogeneity) within a group, or the difference between groups, is great, the "better" or more distinct the clustering. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.

From the foregoing it can be established that the descriptive model recognizes the designs or relationships in data and discovers the properties of the data studied. It does not always forecast to the future as would be seen in the Predictive model.

A. Predictive Model

Time-series methods are also parts of Predictive analytics, making use of methods such as moving averages, exponential smoothing, autoregressive models, linear, non-linear and logistic regression Souza, 2014. "What is a timeseries database?" A time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, utility studies, inventory studies, yield projections, workload projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and medical treatments. A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

V. METHODOLOGIES OF DATA MINING

A. Neural Network

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [23]. This powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in fig.1. Artificial neural network has become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technologies. ANN is an adaptive, nonlinear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the nonlinear characteristics of ANN provide it lots of flexibility to achieve input output map.

B. Decision Trees

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch

represents an outcome of the test and tree leaves represent classes or class distribution. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached [5].Decision tree is represented in figure 2. Fig 2 Decision tree Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. The author defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability [6]. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus marketing managers make use of segmentation of customers, products and sales region for predictive study. These predictive segments derived from the decision tree also come with a description of the characteristics that define the predictive segment. Because of their tree structure and skill to easily generate rules the method is a favoured technique for building understandable models.

C. Genetic Algorithms

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can buia better solution if we somehow combine the "good" parts of other solutions (schemata theory), just like nature does by combining the DNA of living beings [7].

Genetic Algorithm is basically used as a problem-solving strategy in order to provide with a optimal solution. They are the best way to solve the problem for which little is known. They will work well in any search space because they form a very general algorithm. The only thing to be known is what the particular situation is where the solution performs very well, and a genetic algorithm will generate a high-quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problems.

Genetic algorithms (GAs) [8] are based on a biological application; it depends on theory of evolution. When GAs are used for problem solving, the solution has three distinct stages: • The solutions of the problem are encoded into representations that support the necessary variation and selection operations; these representations, are called chromosomes, are as simple as bit strings. • A fitness function judges which solutions are the “best” life forms, that is, most appropriate for the solution of the particular problem. These individuals are favoured in survival and reproduction, thus giving rise to generation. Crossover and mutation produce a new gene individual by recombining features of their parents. Eventually a generation of individuals will be interpreted back to the original problem domain and the fit individual represents the solution.

D. Rule Extraction

The taxonomy of Rule extraction contains three main criteria for evaluation of algorithms: the scope of dependency on the black box and the format of the extract description. The first-dimension concerns with the scope of use of an algorithm either regression or dimension focuses on the extraction algorithm on the underlying black-box: independent algorithms. The third criterion focuses on the obtained rules that might be worthwhile algorithms. Besides this taxonomy the evaluation criteria appear in almost all of these surveys rule; Scalability of the algorithm; consistency.

To remove the deficiency of ANN and decision tree, we suggest rule extraction to produce a tran It is becoming increasingly apparent that the absence of an explanation capability in ANN systems limits the realizations of the full potential of such systems, and it is this precise deficiency that the rule extraction p Experience from the field of expert systems has shown that an explanation capability is a vital function provided by symbolic AI systems. In particular, the ability to generate even limited explanations is absolutely crucial for user acceptance of such systems.

Table 1: Comparison of Various Techniques

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 Kmean Algorithm	Lung cancer Data	9	WEKA	Complexity is high.	Foggy k-mean performs well as compared to Kmeans
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	BF Tree performs well as compared to

					execution time.	other tree classifiers
Qasem A. et.al	Decision tree	e 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	multidimensionality 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	Clusters are created and analyze them with the help of SOM	2	WEKA	Clusters are created and analyze them with the help of SOM	improve the accuracy of analysis compare to kmeans 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.

VI. 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.

From the analysis, the purpose of Predictive model is to determine the future outcome rather than current behaviour. Its output can be categorical or numeric value. For

example, given a prediction model of credit card transactions, the likelihood that a specific transaction is fraudulent can be predicted. Another Predictive model discussed, Regression is a supervised learning technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the development of a model that can predict these ttribute values for new cases. The second model of data mining, Descriptive is normally used to generate frequency, cross tabulation and correlation. Descriptive model can be defined to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data. Summarization as discussed maps data into subsets with associated simple descriptions. It is therefore advisable to use the Predictive model to the Descriptive model since the latter does not forecast into the future. If the conception of computer algorithms being based on the evolutionary of the organism is surprising, the extensiveness with which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new and important area of research and ANN itself is a very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies.

REFERENCES

[1] AbdelghaniBellaachia and ErhanGuven "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, **vol. 6, pp. 234-239,2010.**

[2] 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, pp. 23-128,2010.**

[3] AzharRauf, Mahfooz, Shah Khusro and HumaJaved"Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, **vol. 12, pp. 959-963,2012.**

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

[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 Scientific Research, **vol. 5, pp. 959-9632012.**

[6] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International Multi Conference of Engineers and Computer Scientists, **volume 3, issue 12, pp551-559, IMECS,2009.**

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

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

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

[10] Akhilesh Kumar Yadav, DivyaTomar and SonaliAgarwal Clustering of Lung Cancer Data Using Foggy KMeans", International Conference on Recent Trends in Information Technology (ICRTIT), **vol. 21, pp.121-126.,2014.**

[11] [Sanay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-means Clustering", **vol. 8, pp. 142-147.2014.**

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

[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, pp. 32-38, (2013),

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

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

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

[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, 3rHimani Rani et al, International Journal of Computer Science and Mobile Computing, **Vol.8 Issue., pg. 15-22, 5 may 2019.**

[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.

[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**.

[22] Mortenson, M. J., Doherty, N. F., & Robinson, S. (Operational research from Taylorism to tera bytes:a research agenda for the analytics age. European Journal of OperationalResearch, **583-595,2014**

[23] Delen, D., &Demirkan, H. Data, information and analytics as services. Decision Support Systems, **359- 363.2013.**

[24] .Lechevalier, D., Narayanan, A., &Rachuri, S. Towards a DomainSpecific Framework for PredictiveAnalytics in Manufacturing. IEEE International Conference on Big Data Gaithersburg: National Institute of Standards and Technology, (**pp. 987-995),2014.**

AUTHORS PROFILE

Mr. C. Ganesh is a Full-Time Research scholar in the department of Computer Science, S.V. University College of Commerce Management and Computer science, Tirupati, Andhra Pradesh-India. He is pursuing Ph.D. under the guidance of Dr.E. Kesavulu Reddy in the department of Computer Science, S.V. University College of CM&CS, Tirupati.

Dr.E. Kesavulu Reddy

Department of Computer Science, S. V. University College of CM & CS, Tirupati, Andhra Pradesh-517502, India.

He is working as a Senior Assistant Professor in the Department of Computer Science, Sri Venkateswara University College of Commerce Management and Computer Science, Tirupati, Andhra Pradesh-India. He received master of computer Applications on 2002 from S.V.University, Tirupati, Andhra Pradesh, India. He completed Master of Philosophy in Computer Science on 2006 from Madurai Kamraja University, Madurai, Tamilnadu, and Doctor of Philosophy in Computer Science 2012 from S.V.University, Tirupati, Andhra Pradesh, India. His research interest includes Elliptic Curve Cryptography-Network Security, Data Mining in the Computer science. He had published 50 papers in various International Journals. He had attended and presented 50 papers in various International and National conferences. He had organized two National Conferences i.e. "National Conference on Information Security & Internet of Things (ISIoT-2K19) 20-21, December 2019, and National Conference on Information Security & Data Security in Cloud Computing (ISDSCC2K21) 29-30 April 2021. A PhD student has been awarded under his Supervision and one Submitted during 2014 to 2020. He received Dr. Surveypalli. Radhakrishna Life –Time Achievement National Award with Gold Medal, Memento and Certificate from IRDP Group of Journals, Chennai on 30th May 2018. He was honoured with "Fellow of Computer Science Research Council (FCSRC)" from Open Association of Research Society from Global Journals, U.S.A) on 31st January 2019 for the performance of published research work in the world. He was awarded with "Best Outstanding Researcher 2020" International Award and Best Outstanding Scientists 2020 with Gold Medal, Memento and Certificate from Kamraja Institute of Higher Education Thane, Madurai-Tamilnadu. Also he received "Best Outstanding Scientists 2021" International award from International Scientists on Science, Engineering & Medicine 2021, VDGOODTM TECHNOLOGY FACTORY, Coimbatore, Tamilnadu, India.

