Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 05, May -2018

A LITERATURE REVIEW ON PREDICTION TECHNIQUE IN DATA MINING

Kamini P Parekh

Computer Department, CSPIT, Changa, kimipanchal@gmail.com

Abstract *— In this paper I stated summary of different papers which are based on prediction. Prediction means to predict continuous value of the function. It is used to predict missing or unavailable numerical data values rather than class labels. Prediction also encompasses the identification of distribution trends based on the available data.*

Keywords- *Prediction, Naïve Bayes algorithm, Back propagation, Neural Network, Genetic algorithm*

I. INTRODUCTION

Prediction is similar to classification. First construct a model then use that model to predict unknown values. Major methods for prediction is regression. Mainly there are two types of regression: liner and multiple regressions, non-linear regression. Main difference between prediction and classification is, classification refers to predict categorical class label where prediction refers to continuous valued function.

II. LITERATURE REVIEW

2.1 "Crime analysis and prediction using data minig"-Shiju Sathyadevan, Devan M.S, Surya Gangadharan S. IEEE-2014

Here the paper is based on systematic approach identifying and analyzing patterns and trends in crime. The concept of data mining is used for extracting the information from unstructured data. Step by step procedure follows during the whole work. Data collection, classification, pattern identification, prediction and visualization these all are the steps of proposed work. In data collection the data is collected from the various sources like blogs, news sites, social media, RSS feed etc. Because of unstructured data here Mongo DB is used to store data [1].

Now for classification Naïve Bayes algorithm is used. The algorithm classifies a news article into a crime type to which it fits best. Using the Naïve Bayes algorithm one model is created for training crime data related to different crimes. And for testing the model test data are applied. Test result shows that Naïve Bayes algorithm gives 90% accuracy. Here they also used Name Entity recognition in the crime articles which classify elements of text into predefined categories such as person names, organizations, locations, date time etc. now pattern identification has been done followed by the classification for identify trends and patterns in crime [1].

Here Apriori algorithm is used to determine association rules which highlight general trends in database. As a result of this phase crime pattern for a particular place is known. Now for prediction the decision trees concept is used. Because of it is easy to understand and interpret also it is robust and work well on large data sets. Basically, there are sets of questions to check whether it satisfy the condition or not. If the first condition is satisfied, then they check for the next case. If the first condition itself is not satisfied, then there is no need to check the rest. For visualization heat map is being used here which indicate level of activity. The map shows darker color for low activity and lighter color for high activity.

So, from the map we can prevent crimes by taking preventives mechanisms like night patrolling, fixing burger alarms, fixing CCTV camera etc. so as a conclusion their system can predict the crime prone regions in India on particular day. And by this prediction we can prevent crime by taking necessary actions.

2.2 "Vegetable Price Prediction Using Data Mining Classification Technique"-G.M. Nasira, N. Hemageetha. Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering. March 21-23, 2012

A price prediction model was set up by applying the neural network. They have taken an example of tomato and the parameter of the model are analyze through experiment. Here they use back-propagation algorithm of artificial neural network. Data collection and data preparation: they take vegetable price as experimental data. For creating a model, the price data of tomato from 2009-2011 are taken. Data are collected from the website www.tnau.ac.in of Coimbatore market. For data Normalization minimax normalization is used. The normalize data set will be used to train validate each ANN [2]. The data set will be split into two subsets one for network training and another for network validation.

Configuration of the networks depends on the number of hidden layers, number of neurons in each hidden layers and activation functions.

Three-layer feed forward network structure is used for monthly and weekly vegetable price prediction. Here they have developed code using MATLAB. BPNN is constructed using Jan-2009 to May-2011. Monthly price data and later month's data are used to test model. Experimental results show the absolute error of monthly price prediction is within 10% so the accuracy is up to 90%. They have also predicted weekly price by constructing BPNN using 135 weeks. Here they get absolute error 25% accuracy is up to 75% [2]. In this paper back propagation neural network prediction model of vegetable market is established. The results show that neural network is one way of predicting the market price of the vegetables.

2.3 "Big data and predictive analytics in ERP system for automating decision-making process"- Prof. M. S. Prasada Babu, S. Hanumanth Sastry, IEEE-2014.

In this paper the authors have focused on predictive capabilities of ERP system, to analyze current data and historical facts in order to identify potential risks and opportunities for any organization. Here the author introduced a new implementation framework for Big Data predictive analytics aimed at automating operational decision making in SAP ERP systems [3]. In predictive modeling data is collected, a statically model is formulated, predictions are made, and the model is validating as additional data available from other sources.

With the Big data and predictive analytics, we could combine business knowledge and data mining techniques to achieve insights into business data. In this paper author have presented issue in automation of decision making forecasting processing in ERP system.

2.4 "Genetic Neural network Based Data Mining in Prediction of Heart Disease Using Risk Factors. "Syed Umar Amin, Kavita Agrawal, Dr. Rizwan Beg. – Proceeding of 2013 IEEE conference on information and communication technologies (ICT 2013).

In this paper they presented a technique for prediction of heart disease for using major risk factors. The techniques involve neural network and genetic algorithms. The system was implemented in MATLAB and predicts the risk of heart disease with accuracy of 89% [4]. Researchers have used prediction algorithm in adapted form of simplified score sheets that allow patient to calculate risk of heart disease. The Framingham risk score(FRS) is popular risk prediction criteria which is used in algorithm for heart disease prediction. They have collected data of the 50 people from the surveys done by the American heart association.

Now the data analyses have been carried out in order to transform data into useful form. This system uses back propagation algorithm for learning and training the neural network. A multilayer feed forward network is used having 12 input nodes, 10 hidden nodes, and 2 output nodes. After performing operation on GA and neural network approach the accuracy of prediction heart disease on training data was calculated as 89% and accuracy on validation data was 96.2% [4]. So, the result shows GA and neural approach give better average prediction accuracy then the traditional ANN.

2.5 "Cycle Time Prediction in Wafer Fabrication Line by Applying Data Mining Methods."-Israel Tirkel, IEEE-2011.

In this paper they have developed cycle time prediction models by applying machine learning and data mining methods. Two types of classification are used the best fitted decision tree model and the best neural network model. The dynamic CT prediction model which can be used to predict CT of a single operation step, a line segment or a complete production line. Their Work follows the face of the cross industry standard process for data mining.

The data were taken from fab manufacturing execution system, which contents records of all operational information and transactions of wafer lots processed in the production line. The raw data were extracted on MS Excel spreadsheet for initial analyses and then uploaded to SPSS. After that on verification face incomplete data and inadequate content were eliminated. Redundant data was also eliminated, so the data set was reduced. Now they have applied separation and extraction on the data.

Now the data set form was split into two parts: i) A training data set with 75% of records for the model training. ii) A testing data set with 30% of the records for evaluating the prediction model results. Future set selection process is carried out for their assignment as futures in the prediction models [5]. Three operational sets were generated i) Full-19 features ii) Manual-11 features iii) Auto-7 features. Now the target functions and model selection has been done. The best result of both models was obtained using the full-19 feature set tested with 30% of data set. DT model achieved 76.5% accuracy and the NN model achieved 87.6% accuracy [5].

III. CONCLUSION

By historical figures we can predict the values for forecasting. Also, by prediction we can predict the data in early stage. Prediction is also very useful for crime analysis to identify and analyzing patterns and trends in crime. We can also predict and diagnosis a various disease with good accuracy. Prediction is also helpful to farmer and government to make effective decision based on price prediction. It can identify potential risk and opportunities for any organization.

REFERENCES

[1] Shiju Sathyadevan, Devan M. S., Surya Gangadharam S., “Crime analysis and prediction using data mining” IEEE-2014.

[2] G. M. Nasira, N. hemageetha, “Vegetable Price Prediction Using Data Mining Classification Technique” Proceedings of the international Conference on Pattern Recognition, Informatics and Medical Engineering. March 21-23. 2012.

[3] Prof. M. S. Prasada Babu, S. Hanumanth Sastry, “Big Data and Predictive Analytics in ERP System for Automating Decision making Process”, IEEE-2014.

[4] Syed UmarAmin., Kavita Agrawal. Dr. Rizwan Beg, “Genetic Neural Network Based Data Mining in Prediction of Heart Disease using Risk Factors.”, Proceeding of 2013 IEEE conference on information and communication technologies (ICT 2013).

[5] Israel Tirkel, “Cycle Time Prediction in Wafer Fabrication Line by Applying Data Mining Methods.”, IEEE-2011.

[6] Jiawei Han, Micheline Kamber and Jian Pei “Data Mining: Concept and Techniques”, The Morgan Kaufmann series in Data Management System. 3rt ed.