Enhanced Crisis Management: Predictive Strategy for Human Blood Group and Organ Demand using Polynomial Random Forest Algorithm

Karthik Elangovan, Sethukarasi.T

ABSTRACT---Prediction is one of the important tasks in the Machine learning. It is the emerging trend in the Data mining as an era of internet the data has evolved into big data in the concern of volume, velocity, veracity and verity. There is a tremendous growth in the prediction techniques thereby a huge volume of data has been used and processing speed has increased through In-memory analytics. Prediction attempts to identify a new pattern that helps to predict the future events. Structured data are being collected from the various resources such as web content, social networks, sensors and made available across different domain. This paper focuses on predicting the events which make crisis in future, as of to whether they may happen or not. Proposed algorithm is Polynomial Random Forest Prediction which finds out the expected probability for the event to happen in terms of percentage. This algorithm takes a cluster of attributes as input from a data source that may depict the present condition along with an interrogative attributes that the user may want to know by the predictive capability of the algorithm. The percentage is further calculated using polynomial equations, data mining, and random forest. Hence empirically we could calculate the probability of occurrence of the event in future. This predictive model can be used for various applications amongst which we focus on human blood group and organ demands.

Keywords: Data mining, Machine learning, Percentage, Probability, Prediction, Random forest.

1. INTRODUCTION

A prediction is often a forecast about an event that is totally uncertain to happen in future. It is done with the help of existing knowledge and past experiences of that event. There are a lot of prediction algorithms that have been devised in mere present but most of them often fail to have a very high success rate. Predictions are tough to be determined especially about the natural events that are yet to occur. Machine learning [10] is an effective tool in making forecasts and calculations based upon the prevailing data.the concept of data mininghave a process of computation that determines patterns in large data sets involving methods of machine learning tools along with knowledge base systems.With this we can extract information from the past experiences and turn them into an understandable structure for further use. So hence this data mining[9] becomes a powerful tool in extracting information from the database and makes the existing systems think like a human being.

Revised Manuscript Received on February 14, 2019.

KarthikElangovan, (Assistant Professor), Computer Science and Engineering, S.A. Engineering College, Thiruverkadu, Tamil Nadu, India.

Dr.Sethukarasi.T,(Professor and Head), Computer Science and Engineering, R.M.K Engineering College, Gummidipoondi, Tamil Nadu, India.

2. LITERATURE SURVEY

Share market is one of the important fields to predict the next event that can be done by ANN(Artificial Neural Network) [1]. Computing the prediction through ANN contains two phases. First phase is the normalization of the input,which makes the input to process efficiently in the forthcoming modules to achieve accuracy in every phase of the output. The second phase is the prediction where the ANN technique contains input layers, hidden layers and output layer. Hidden layers neurons generate the final output which is compared with the real output and calculate anerror. The error is generated from the Propagation Phase which is used to update the weight in the hidden layers.

Logistic regression, Decision tree(CART) and ANN [2]are the best approaches to detect the fraud in the financial transition and this proposed method gives the classification rate for ANN and CART for sample data sets Logistic regression model is used to identify the exact features from the large set of features, the error is ascertained at the yield and disseminated back through the network layers. In this strategy artificial neural network is trained thereby squared error is minimized between the actual value and predicted value. The value that has been predicted has an associated error which indicates how close the predicted value comes to the observed value. The standard error estimation is given by,

$$S_{ee} = \sigma_y\sqrt{1 - r^2}$$

Where, the Pearson-product moment and the standard deviation of the Y-variable (SDY) between the X and Y variables is r.Classification and regression tree are non-parametric statistical models used to determine the response variables from the exhaustive variables. This is the only available procedure for analysing both categorical data and continuous data. Li Jun and Zhang Peng [3] proposed Probable customers can be classified from the large number of internet handlers through the data mining models. Advertisement can be delivered only to the potential buyers not all the users. This greatly reduces the advertising expense. Feature selection has concentrated only on search keywords and web links. Semantic issues of the keywords are not addressed. It is discussed only on structured Data but not applied for Big Data.

Xindong Wu, et al. [4] have proposed how the data mining concepts are helpful to implement in the Big data processing. HACE theorem is used for characterizing the

features of the Big Data revolution. It considers Big Data processing model from the data mining perspective. Security and privacy for big data issues are addressed; this model requires relationship between sample models and data sources.Bryan Higgs and Montasir Abbas [5] have suggested a two-step algorithm and used for the segmentation and clustering of vehicle driving behaviours. Unknown driving states are predicted from known driving states using several parameters of car driving. This model does not provide individual model for each driver. Only 30 clusters have been used to model the driving pattern.

ArshdeepBahga and Vijay K. Madisetti [6] have conferred hybrid approach that can analyse sensor data in existent time and by using the global information on previous errors the new errors were forecastfrom a huge number of machines which utilisesthe cloud-based case-based reasoning (CBR).CBR finds solutions based on past experience. This past experience is organized and represented as cases in a case base.

The implementation of Map Reduce and Pig Latin can examine and update enormous machine data for case-construct creation on a timespan of seconds in the cloud instead of in minutes.Wenping Zhang and Raymond Lau [7] have proposed a hybrid symbolic and quantitative approach used to improve the recommender agents to provide prediction effectiveness, learning autonomy and explanatory power further it applies only for specificdomain. In these papers, a list of various algorithms such as linear regression, logistic regression,naive Bayes, k-nearest neighbour, decision trees, random forest, adaboost, artificial neural networks and support vector regression algorithms have been presented for predicting new patterns or events.

3. TECHNIQUES FOR PREDICTION

A) Decision trees:

Decision trees are a class of data mining techniques that have roots in traditional statistical disciplines such as linear regression [12]. Decision treesalso share roots in the same field of cognitive science that produced neural networks [15]. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree [11] consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes).

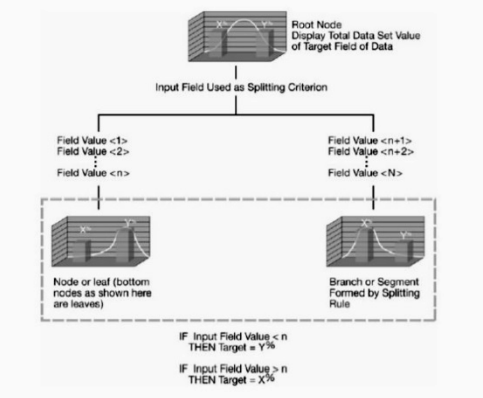

Fig 1.The decision tree

B) Random Forest Algorithm:

Random forest algorithm is a supervised classification algorithm that creates the forest with a number of classification trees. The input attributes are fed to each of the trees in the forest that yield the classification of data. Each tree gives a classification with high information gain for a particular class. The forest chooses the classification having the high information gain among all the trees in the forest. The forest becomes robust based on the number of trees in it which produces high accurate classification [18]. Random forest (RF) prediction algorithm is used to predict the missing data as follows:-

RF Prediction $s = \frac{1}{K} \sum_{k-1}^{k} K^{th}$ tree response

where, k = 1, 2, … , n.

Algorithm for Random Forest:

1. Select "k" random features from "M"total features, where$1 \leq k < M$.
2. Determine the node distance "d" among the "k" features by the best divider point.
3. Split the node into child nodes using the best divider.
4. Iterate steps 1 to 3 till unit node attainment.
5. Construct the forest by iterating the steps 1 to 4 for multiple times to create several trees.

Each tree is grown as follows:-

1. Train the 'N' sample cases based on the original data at random - but *with replacement*. This sample will be treated as the training data set for testing the data.
2. Among "M" input attributes, "k" is an integer which is randomly selected and the best partition on these "k" is used to divide the node. The value of "k"remains constant during the growth of the forest.
3. Each tree is expanded to the maximum level possible. There is no pruning.

Random forest algorithm[19] runs efficiently on large data sets and calibrates the important variable for classification. It mitigates the generalization error during the building of random forest. Missing data can be accurately predicted. It determines clustering and locating outliers of data. This algorithm can be extended for unlabelled data using unsupervised clustering. Random forest does not over fit.

C) Generalized linear regression:

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression[12] that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Ordinary linear regression predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors). This implies that a constant change in a predictor leads to a constant change in the response variable. This is appropriate when the response variable has a normal distribution.

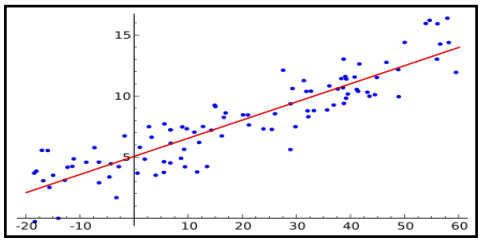

Fig 2. Generalized linear regression

4. RESULTS & DISCUSSIONS

Our proposed algorithm works in two separate threads simultaneously running at the same time. The final prediction [17] gives us what organs and equipment will be needed to treat the patient.

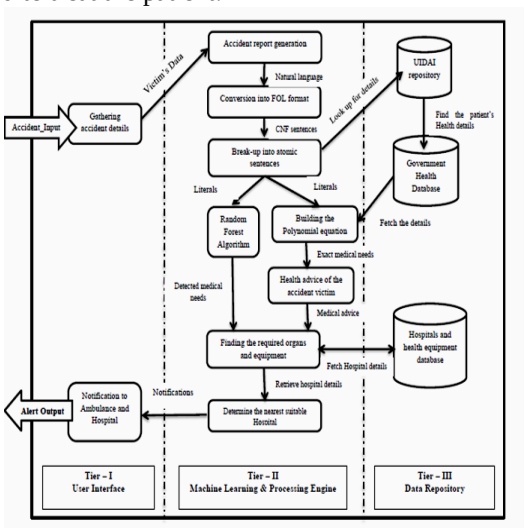

Fig 3.Architecture Diagram for proposed system

The accident reporter is intended to give a set of statements which determine the details of the accident and the victim. This needs to be given to the system as an input to the system which will determine and predict the output for the organs [8] that will be needed and equipment and facilities required to treat the victim based on the given statements along with the previous data that is obtained from data mining. The given set of statements will be converted into predicates. From here the process is divided into two parts. First part is related to the accident and the next depends on the victim.

The first threat uses the predicates related to the accident and the present conditions to build a Decision Tree [12]. The details of how an accident takes place and what will be the consequences of the accident are given to the system well in advance. The system compares each predicates with the predefined conditions to build the decision tree. Using the help of the decision tree we find out all the parts and organs that will be needed to treat the patient (victim).

The second process depends upon the victim. The health conditions of the victim are considered to be stored in the hospital database that willbemaintained by the government. This idea employs the concept of Aadhar card. We consider the patient to have an Aadhar card registered. His/her biometric [13] identities can be used to find out the condition of the victim through the database. Even if the biometrics is not possibly helpful, if the victim carries the Aadhar card, it could be used to find the necessary details.

Using the victim's health conditions, it is possible to find out which organs will be needed or which parts of the body will need special attention or handled safely. Consider a heart patient. Even if the accident is not related to heart, the patients need special equipment to treat them. So a polynomial regression is done on the details of the patient along with the type of accident to find the required doctors, equipment and hospitals.

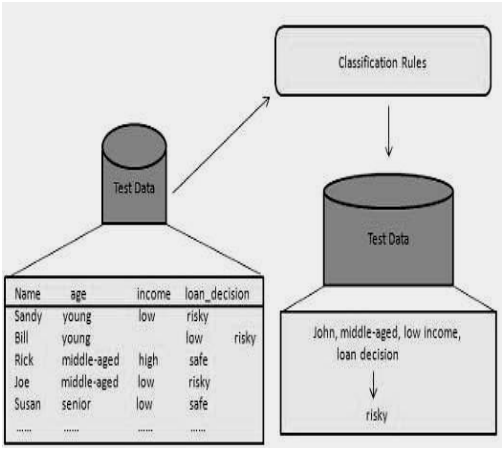

Fig 4.Data mining to get the factors of the test data

From the data, that is analysed and obtained from the data mining process [16], for each condition of the victim, a polynomial equation is built with the factor that is found using the type of accident. Consider the below equation where y is the requirement to treat the victim and x is the condition of the victim.

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \ldots\ldots + a_m x_i^m + \varepsilon_i$$

where i = 1, 2,.., n.

The above equation is a type of polynomial equation used in polynomial prediction. The coefficients of x and the power of x are determined by the type of accident that has occurred. By solving the equation we get to know whether the victim's weak parts need to be treated or replaced due to the accident. The equipment needed for treatment can also be found once the requirements and special care details are found. Upon finding the needs for treatment, we need to find which hospital has the required doctors and equipment. For this case, we use a Random Forest algorithm[14].This builds a set of decision tree for each requirement.The outcome of each tree is used to rate the most successful hospital to handle such cases of victim.

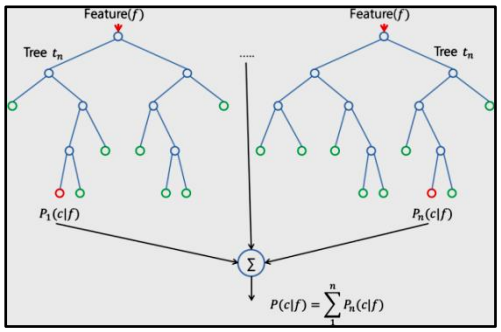

Fig 5. Random Forest

The hospital must have the required doctors and equipment to treat the most affected part of the victim due to the accident based on the victim's condition. If the hospital meets such needs and does not support the equipment to treat the other needs or employs doctors for those needs, the system notifies the hospital that is good at handling such cases to send the doctors and equipment to treat the victim who is admitted in another hospital.

Eventually, the system notifies all the hospitals that are to work in the case and the ambulance suited to take the patient to the hospital safely on time. The ambulances are selected based on the current location and facilities that meet the requirements.

5. CONCLUSION

Hence we have designed an algorithm for the predictions in crisis management. The proposed algorithm can be implemented only when the Aadhar card is effectively used by the citizen. An half of this algorithm can be used to predict which parts of the victim need to be tested as soon as the medic meets the victim so that the requirements are made more accurate and the treatment is made effective and successful. The treatment of the victim is considered as the main objective and the hospitals are required to cooperate with the system and the government to help the victim receive a quality treatment alongside the absolute treatments needed.However the Random Forest Algorithm lacks the tree-pruning which may incredibly affects memory space used and time spent for searching in the forest. This can be enhanced by using un-supervised learning which handles unlabelled data such as the patient health details does not possess the required labels for data that proportionally upgrades the overall system performance.

REFERENCES

1. ZabirHaiderKhan,TasnimSharminAlin,Md, May 2011 AkterHussain,Price Prediction of Share Market using Artificial Neural Network (ANN), International Journal of Computer Applications (0975 – 8887),22(2).
2. Chi-Chen Lin, An-An Chiu, Shaio Yan Huang, David C.Yen,2015. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts judgments, Knowledge-Based Systems ,89:459–470.
3. Li Jun, Zhang Peng,2013.Mining Explainable User Interests from Scalable User Behavior Data, Procedia Computer Science,17: 789 – 796.
4. Xindong Wu, XingquanZhu,Gong-Qing Wu, and Wei Ding,2014.Data Mining with Big Data, IEEE Trans on Knowledge and Data Engineering,26(1): 97-106.
5. Bryan Higgs and MontasirAbbas,2015. Segmentation and Clustering of Car-Following Behavior: Recognition of Driving Patterns IEEE Trans On Intelligent Transportation Systems,16: 81-90.
6. ArshdeepBahga and Vijay K. Madisetti, October 2012. Fellow, IEEE Analyzing Massive Machine Maintenance Data in a Computing CloudIEEE Transactions On Parallel And Distributed Systems,23(10).
7. Wenping Zhang and Raymond Lau, 2012. Mining Contextual Knowledge for Context-Aware Recommender Systems, Ninth IEEE International Conference on e-Business Engineering, 355-360.
8. D.B. Neill and G.F. Cooper, "A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization,"Machine Learning, vol. 79, 2010, pp. 26
9. T. Fawcett and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behaviour," Proc. 5th Int'l Conf. Knowledge Discovery and Data Mining, 1999, pp. 53–62. 1–282.
10. D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, 2003, pp. 993–1022.
11. Biau,G. Analysis of a random forests model. Journal of Machine Learning Research, 13:1063–1095, 2012.
12. Breiman, L., Friedman, J., Stone, C., and Olshen, R. Classification and Regression Trees. CRC Press LLC, 1984.
13. Devroye, L., Gy¨orfi, L., andLugosi, G. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, 1996.
14. Xiong, C., Johnson, D., Xu, R., and Corso, J. J. Random forests for metric learning with implicit pair wise position dependence. In ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 958–966, 2012.
15. R. J. Van Eyden, "Application of Neural Networks in the Forecasting of Share Prices" Finance and Technology Publishing, 1996.
16. Chiu, S. L.; 1994, "Fuzzy model identification based on cluster estimation", Journal of Intelligent and Fuzzy Systems, 2, John Wiley & Sons, pp. 267-278.
17. Justin Wolfers, Eric Zitzewitz, "Prediction markets in theory and practice", national bureau of economic research, pp.1-11, March 2006.
18. Arnu Pretorius, SuretteBierman and Sarel J. Steel, 2016, A Meta-Analysis of Research in Random Forests for Classification, International Conference (PRASA-RobMech), 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics
19. Alejandro González, Antonio M. López, 2017, On-Board Object Detection: Multicue, Multimodal, and Multiview Random Forest of Local Experts, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 47, NO. 11, pp 3980-3990