IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 2, Ver. IV (Mar.-Apr. 2017), PP 19-24
www.iosrjournals.org

Score Level Fusion Based Death Prediction using Data Mining Techniques

Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, Ajit Danti, S.N. Bharat
NES Research Foundation, JNN College of Engineering, Shimoga, Karnataka, India

Abstract: *This paper presents the study regarding the analysis of death prediction using data mining techniques. In this paper four different supervisor machine learning algorithmsare considered for mortality rate prediction of death. Further score level fusion is employed for optimum decision by various combinations of classifiers for the prediction of death. Score level fusion is robust enough to predict the death. The proposed model is evaluated by considering publically available Queensland government dataset. Theresults of the proposed model reveal interesting facts with prediction efficiency.*

Keywords:*Web mining, mortality prediction, Machine Learning., fusion, score*

I. Introduction

Web data mining is one of the allied themes of data mining technique, which is used to extract the information from the webpage for analyzing the facts for specific applications. Web data includes web documents, web pages, hyperlinks between web pages and finally log information on web sites. This paper addresses two different problems like predicting and analyzing the death using web data mining technique. Generally prediction is identification of one thing entirely based on the description of the related thing. Mathematically this can be described as predicting $\boldsymbol{i+1}^{th}$ case with the help of first $\boldsymbol{i}$ cases by considering the data available in the web site. The website will be having the statistics of the death rate between x and y years are because of different reasons and predicting algorithm will predict death rate for the year z. Based on the data mining techniques, the death rate of the human beings is predicted. The prediction is accomplished by set of machine learning techniques which consists different stages like training and testing. Since machine learning algorithms are considered for predicting the death rate, previous year's statistics behaviors like a training data. In this paper four different supervised learning algorithms are considered for predicting the death rate like interval valued classifier, nearest neighbor classifier, centroid classifier and decision tree classifiers for mortality rate prediction using score level fusion.

The rest of the paper is organized as follows. In section II a brief literature survey on the exiting state of the art techniques is presented. In section III proposed model for predicting mortality rate of deathis described. Section IV discusses about experimentation and comparative analysis. Paper will be concluded in section V.

II. Related Work

In general forecasting and prediction is the future behavior or future event of the selected data set. The predicting knowledge from the analyzed data is used to predict future behaviors and event. It helps in various domains such as future marketing campaigns, Future event prediction, and pre fetching web pages for improving performance allocating or de-allocating resources and coaching. There are a specific number of researches have been done on Web site related forecasting. There are few number of researches have been done on future event related forecasting.

Enke, D.S.Thawornwong explained the techniques and method of data mining and neural network needed for prediction and forecasting of stock market prices and values. It has been accepted widely by many studies that nonlinearity exists in the financial markets and that neural networks can be used effectively to uncover this relationship [2].H. A. Ahmed Aqlanet. al., explained prediction and the causes of death event is done using web mining techniques. In thisapproach real-time predictions about the likelihoods of future death and disease events of interest [5].

Shoiab Ahmedand A. Danti have explored data mining techniques on rule based classifier using precisionmethods [15]. Gouda. et al. made comparison among the different classifiers such as decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer and used fusion at classification level between these classifiers to get the most suitable multi-classifier approach for each data set[4].K.K.Sureshkumar has used Weka tool to get more accurate stock prediction price and compared with weka classifier methods such as Gaussian processes, isotonic regression, least mean square, and linear regression, multilayer perceptron, pace regression, simple linear regression and SMORegression [7].SolankiA.V. have exploreddata mining technique for classification of sickle cell disease prevalent in Gujarat, From there experimentation it can be inferred that Random tree is better algorithm as it produces more depth decisions respect to J48 for sickle cell diseases[16]. S.C. Dangareand Sulabha S. explained about analyzed prediction systems for Heart disease using more number of input attributes [11].ShantakumarB.Patil,and Y.S.Kumaraswamyhave proposed theoretical underpinnings of the Bayesian approach for classification [14]. V. Krishnaiah et al explored for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient[17].GeramiFarzad et al Predicted and forecasting Workplace Accidents by WEKA Software tool using the linear regression method [3]. VrushaliBhuyar has used classification Techniques on Soil Data and Predicted and forecast Fertility Rate for Aurangabad District [20]. Velide Phanikumar and Lakshmi Velide discussed, processing and predicted nitrogen, phosphorus and sculpture in soil in less time by the linear regression method [**18**].

Vijayarani S. and Sudha S. have compared the analysis of classification function techniques for heart disease prediction [19]. Nan Gao et aldiscussed the idea about Forecasting Model on Emergency Incidents in a city using WEKA software tool [**9**]. Elia Georgiana Dragomir predicted an Air Quality Index forecasting using K-Nearest Neighbor Technique [**1**]. Rajesh Kumar explained about the decision tree method in forecasting the dependent variables like fog and rain for weather forecasting using WEKA [**10**]. S.Dhamodharan used Bayesian classification technique, which is one of the major classification models. The primary goal is to predict the class type from classes such as ‘Liver Cancer’, ’Cirrhosis’, ‘Hepatitis’ and ‘No Disease’ [12].

S.N. Bharath and A. Danti illustrates combined approaches for text classification system. Integer representation is achieved using ASCII values of the each integer and later linear regression is applied for efficient classification of text documents. An extensive experimentation using nearest neighbor supervised learning algorithms on four publically available corpuses are carried out to reveal the efficiency of the proposed technique [13]. Haizhou DU suggesting an idea about Wind Power Load Forecasting based on the data mining classification techniques using WEKA [**6**].

III. Proposed Model

This paper presents supervised learning algorithm for the human death rate prediction. The proposed model is developed by considering five different diseases Trachea, bronchus and lung, Melanoma of skin, Breast, Female genital organs and Male genital organs. The proposed model consists of different stages like visualization of the data, training stage and testing stage as shown in Figure 1.

Visualization Stage:Data visualization stage is considered as important stage of the proposed models. The aim of the visualization stage is to understand the data for mortality rate predictionand analyze the ratio between, death rate with respect to particular disease to particular year. Let Ri be death rate of dieses D_i with respect to the year Y_i. The ratio between R_{i+1}, R_i of D_i and Y_i is calculated. This analysis provides the increase or decrease in rate of the dieses Di. This information is considered for the training of the learning algorithm.

Testing Stage: Once the data visualization stage is completed, the information about mortality rate with respected to different diseases for the year Y_i are considered for training the learning algorithm. Once the training stage is completed, the learning algorithm is considered for predicting the mortality rate for the year Y_{i+n}.The overall procedure of the proposed model is diagrammatically presented in Figure-1. The proposed model is made to work on four different classifier like interval valued classifier, nearest neighbor classifier, centroid classifier and decision tree classifiers. Among the four classifiers, interval valued classifieris specially designed for classification by considering minimumf_i^- and maximum f_i^+ values of the available features. But in this article, minimum and maximum values are estimated based on the ratio between R_{i+1} and R_i.

Algorithm – 1: Prediction of Mortality rate using supervised learning algorithm.
Input: Statistics of mortality rates of five diseases over the years $y_{(i,j)}$.
Output: Predicted mortality rate $mrb_{(i,j)}$.of the diseases
Step1: Let $y_{(i,j)}$: Mortality rate of a disease *i* and for a year *j*.
Step2: Mortality Ratio Base $mrb_{(i,j)} = r_{(i,\,j)} / r_{(i,\,j+1)}$
Where, r(i,j),r(i,,j+1) : Mortality rate of disease for two consecutive years respectively.
Step3: *p1* = Interval _valued_ classifier_$mrb_{\,(i,j)}$
p2 = Nearest neighbor_$mrb_{(i,j)}$
p3 = Centroid classifier_$mrb_{(i,j)}$
p4 = Decision tree_$mrb_{(i,j)}$
where, *p1,p2,p3,p4* are predictions from each classifiers
Step4: Score level Fusion Approach 1

Score s1=Max(p1,p2,p3,p4)
*Step5:*Score level Fusion Approach 2

$$\text{Score } s2 = \frac{2 \times N^{FF}}{N^{TF} \times N^{FT} + 2 \times N^{FF}}$$

Where
N^{TT} *: correct classification from both classifiers.*
N^{TF}*: first classifier classified correctly and second classifier classified wrongly.*
N^{FT}*: first classifier classified wrongly and second classifier classified correctly.*
N^{FF}*: incorrect classification from both classifiers.*

Step6: Fusion   f = Max (s1, s2)
Where   f: fusion indicate final Mortality rate Prediction
End.

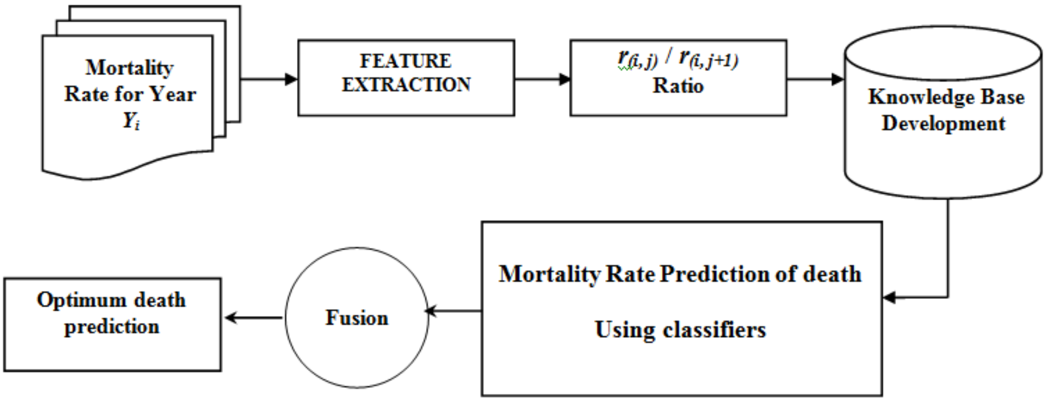
Mortality Rate for Year Y_i
FEATURE EXTRACTION
$r_{(i,j)}$ / $r_{(i,j+1)}$ Ratio
Knowledge Base Development
Mortality Rate Prediction of death Using classifiers
Fusion
Optimum death prediction

Figure 1: Block diagram of Proposed Mortality Rate Prediction of Death

Prediction result obtained by four classifiers and final prediction is determined using score level fusion obtained by two approaches as given below.

Approach – 1:Score Level Fusion Based Approach
To demonstrate the efficiency of the proposed algorithms, score level fusion method is adapted for the four different supervised learning algorithms considered in the paper. Generally voting methods consider the *f*-measure values of the four different learning algorithms and predict the result which is selected by most classifiers.Simple fusion method is formulated as follows.

Classification Score = Max(p1, p2, p3, p4)

Where p1, p2, p3 and p4 prediction result of classifier.
But for some applications, combination of classification algorithms plays a major role in assessing the performance of the model.

Approach – 2:Prediction by Combined Classifier:
In this article four different numbers of classifiers are considered for mortality rate prediction. But to assess the overall performance of the proposed model classifiers will be selected based on few parameters. To achieve the good performance of the proposed model, the performance of the each individual classifier need to be optimized. If more than oneclassifier with minimum marginal performance are being considered, it will be difficult to expect the high level of accuracy in the system. This can be addressed by selecting computationally less expensive and high performance classification algorithm. After the experimentation, another type of confusion matrix *Cm* is generated to calculate the classifier correlation. The confusion matrix *Cm* lists true classes *c* verses the estimated class *ĉ*. This is because of all classes can be enumerated, it is possible to obtain information not only about the correctly classified states N^{TT} and N^{FF}, but also false positive N^{FT} and N^{TF}.

Table 1: Confusion matrix

	True	False
True False	N^{TT} N^{FT}	N^{TF} N^{FF}

Table1 is typical two – class confusion matrix M, where off-diagonal entries present the correlation degree of the two classifier.
Where
N^{TT} *: correct classification from both classifiers.*
N^{TF}*: first classifier classified correctly and second classifier classified wrongly.*
N^{FT}*: first classifier classified wrongly and second classifier classified correctly.*
N^{FF}*: incorrect classification from both classifiers.*

Petrakos et al. [8] presents a classifier correlation analysisS_2 for two classifiers as follows.

$$S2 = \frac{2 \times N^{FF}}{N^{TF} \times N^{FT} + 2 \times N^{FF}}$$

S_2 plays a major role in selecting combination of classifiers for classification fusion algorithm for mortality rate prediction. Result of the selection techniques of classification algorithm based on the petrakos technique is graphically represented in Figure2.

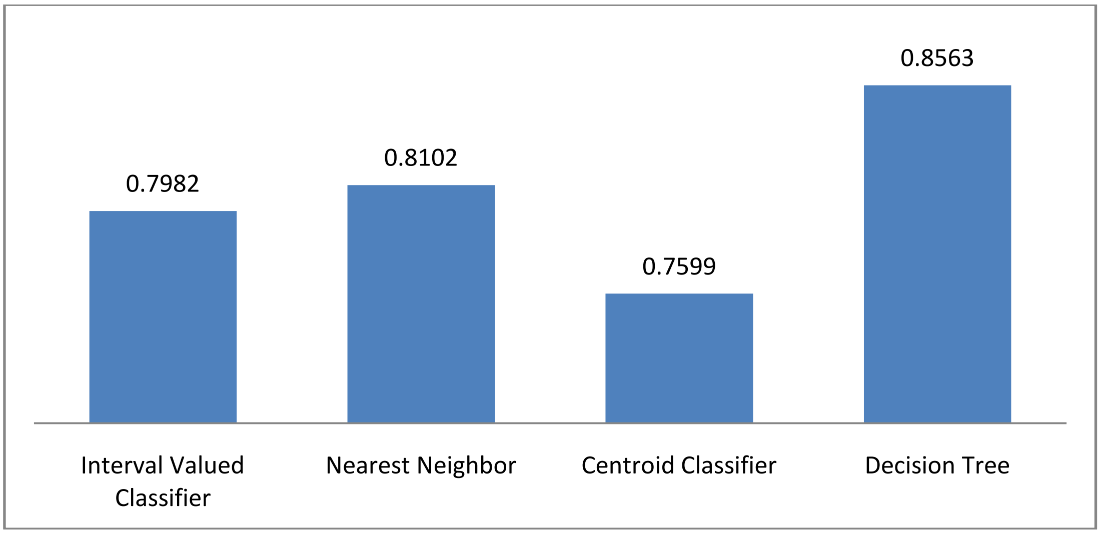

Figure 2: Efficiency of the Individual Learning Algorithms.

IV. Experimental Results

Any systems need to evaluate by considering state of the art publically available datasets. In this article Queensland Governmentdataset is considered for the evaluation purpose. Queensland Government dataset consists of morality rate of the five important diseases viz Trachea, bronchus and lung, Melanoma of skin, Breast, Female genital organs and Male genital organs. The proposed model is made to work on four different classifier like interval valued classifier, nearest neighbor, centroid classifier and decision tree. The datasets consistsdeath data for the year 2011, 2012 and 2013. Table -2 presents the results of the experiments conducted for the evaluation of the proposed algorithms.

Table 2 : Classification results of the proposed model

Classifier Name		Accuracy
p1	Interval Valued Classifier	0.7982
p2	Nearest Neighbor	0.8102
p3	Centroid Classifier	0.7599
p4	Decision Tree	0.8563

It is clear from the Table 2, that decision tree learning algorithm perform well compared to other techniques, due to its ability to capture the knowledge from the training samples.
Table 3 shows the prediction result of combined classifier and Figure 3 shows plot of prediction result by combination classifier.

Table 3: Prediction by Combined classifier

Combination	AccuracyPrediction result
p1-p2	0.8042
p1-p3	0.8042
p1-p4	0.8042
p2-p3	0.8042
p2-p4	0.83325
p3-p4	0.8081

The proposed algorithm is implemented on core i3, 2.66 GHz Processer, 2GB RAM and on Windows 7 platform using MAT Lab version R2012a. Proposed algorithm is computationally less expensive and fusion results are obtained by different combination of classifiers as show in Figure 3.

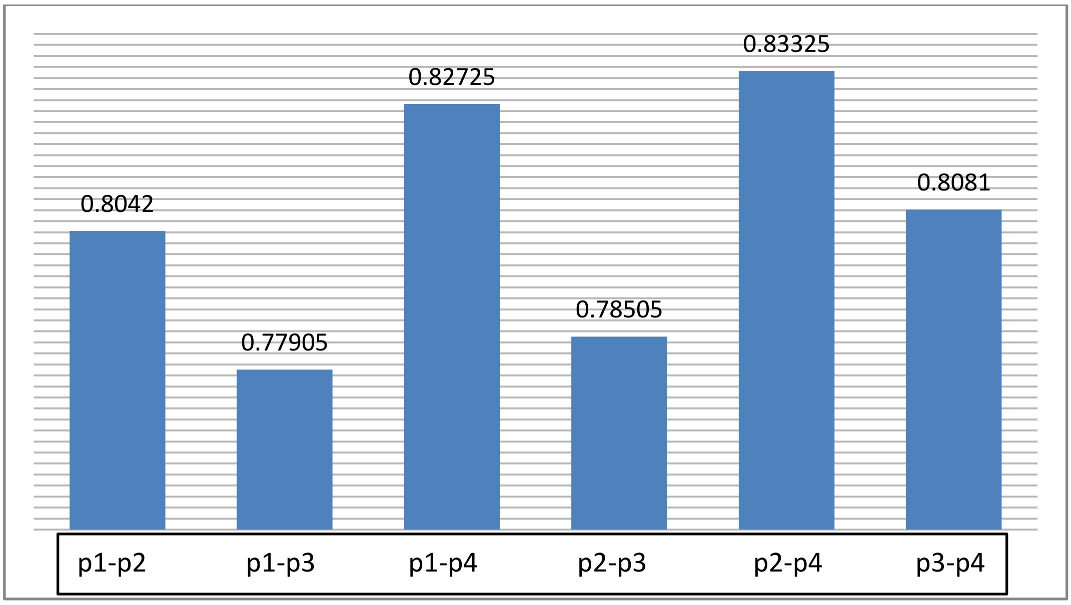

Figure 3: Prediction by combined classifier.

V. Conclusion

The proposed model presents the problem like predicting and analyzing the death using web data mining technique. Four different machine learning algorithm is considered for mortality rate prediction. The proposed model is evaluated by considering publically availablenormal size dataset. The experimental results reveal that, proposed model is provided good results fromfour different classifiers for mortality rate prediction. Further simple fusion technique is applied for analyzing the efficiency of the algorithms. A combination rule to find out best combinations of classifiers is also devised. In the future one can think of designing a generalized combination rule for selecting combination of *n* classifiers for efficient modeling.Evaluation of results and graphical analysis reveal interesting facts that p2-p4 i.e., decision tree algorithms and nearest neighbor pair are better suited for forecasting since these algorithm utilizes complete training sets and produces higher prediction rate, whereas p1-p3 i.e., interval valued classifier and centroid classifier gives poor results due to the fact that it utilizes subset of the training sample

References

[1]. Elia Georgiana Dragomir Air Quality Index Prediction using K-Nearest Neighbor Technique", buletinuluniversităţii petrol–gaze din Ploiesti, Vol. LXII No. 1/2010.

[2]. Enke, D., Thawornwong, S. (2005) "The use of data mining and neural networks for forecasting Stock market returns", Expert Systems with Applications, 29, pp. 927-940

[3]. GeramiFarzad, BartashakMasoumeh, Kourosh Rocky, RaziehHonarmand" Prediction of Workplace Accidents with Knowledge Discovery Approach Using Wekasoftware", Nova Explore Publications, Nova Journal of Engineering and Applied Sciences Vol 2(5), May 2014:1-8.

[4]. Gouda .Salama, M. et al. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of Computer and Information Technology (2277 – 0764), September 2012.

[5]. H. A. Ahmed Aqlan, Shoiab Ahmed, Ajit Danti "Death Prediction and Analysis Using Web Mining Techniques".2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Jan. 06 – 07, 2017, Coimbatore, INDIA.

[6]. Haizhou DU "Intelligent Optimization Research of Wind Power Load Forecasting", Journal of Pattern Recognition & Image Processing 4:4 (2013) 507-513.

[7]. Information Technologies, Vol. 4 (1), 39–45, (2013).

[8]. K.K.Sureshkumar "An Efficient Approach to Forecast Indian Stock Market Price and their Performance Analysis", IJCTA (2011).

[9]. M. Petrakos, I. Kannelopoulos, J. Benediktsson, and M.Pesaresi, The effect of correlation on the accuracy of the combined classifier in decision level fusion, Proc. IEEE 2000 Intl. Geo-science and Remote Sensing Symp., Vol.6, 2000.

[10]. Nan GAO, XuemingShu, JitingXu, Biao Wen, Peng Chen, and PengWu"The Study ofQuantitative Forecasting Model on City Emergency Incidents", International Journal of Information and Education Technology, Vol. 3, No. 5, October2013.

[11]. Rajesh Kumar "Decision Tree for the Weather Forecasting", International Journal ofComputer Applications (0975–8887) Volume 76–No.2, August 2013.

[12]. S.C. Dangare,Sulabha S, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques",International Journal of Computer Applications (0975 – 888), June 2012.

[13]. S.Dhamodharan-Liver Disease Prediction Using Bayesian Classification; Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014.

[14]. S.N. Bharath B, A. Danti. "Document Vector Space Representation Model for Automatic Text Classification". In Proceedings of International Conference on Multimedia Processing, Communication and Information Technology, Shimoga. pp. 338-3442016.

[15]. ShantakumarB.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 , 2009.

[16]. Shoiab Ahmed, Ajit Danti – Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers; International Conference on Computational Intelligence in Data Mining, Volume 1, Pages 171-179, SPRINGER publications, December 2015.

[17]. SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.

[18]. V. Krishnaiah et al "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Scienceand

[19]. Velide Phanikumar and Lakshmi Velide "Data mining plays a key role in soil data analysisOf Warangal region", International Journal of Scientific and Research Publications, Volume 4, Issue 3, March 2014.

[20]. Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

[21]. VrushaliBhuyar "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility y Rate for Aurangabad District", International Journal of Emerging Trends &Technology in Computer Science (IJETTCS) Volume 3, Issue 2, March April 2014.