ISSN 2278-3091

Volume 6, No.5, September -October 2017

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01652017.pdf

ENSEMBLE ONLINE SEQUENTIAL EXTREME LEARNING MACHINE AND SWARM INTELLIGENT BASED FEATURE SELECTION FOR CLEVELAND HEART DISEASE PREDICTION SYSTEM

A.V. Senthil Kumar

Professor,Department of Post Graduate and Research in Computer Applications, Hindusthan College of Arts and Science,
Coimbatore- 641 028, Tamil Nadu, India.
avsenthilkumar@yahoo.com

ABSTRACT: In Healthcare industries generally clinical diagnosis is done mostly by doctor's expertise and experience. Computer Aided Decision Support System plays a major role in medical field. With the growing research on heart disease predicting system, it has become important to categories the research outcomes and provides readers with an overview of the existing heart disease prediction techniques in each category. The purpose of this paper is to develop a cost effective treatment using data mining Ensemble of Online Sequential Extreme Learning Machine (EOS-ELM) for facilitating data base decision support system. Heart disease patient database is collected from Cleveland Heart Disease Dataset (CHDD) available on the University of California, Irvine (UCI) Repository. Since the database samples consists of huge attribute and selection of best attribute from this CHDD becomes very important for prediction accuracy. So the dimensionality reduction is done by using Modified Genetic Algorithm (MGA) from these features set significantly speeds up the prediction task. This research paper added two more attributes i.e. obesity and smoking. The results from experiments returned with diminishing fact that there is considerable improvement in classification and prediction. The proposed EOS-ELM works as promising tool for prediction of heart disease when compared to other data mining classification techniques, namely Naïve Bayes (NB), Decision Tree (DT) and Artificial Neural Networks (ANN) and are analyzed on Cleveland Heart Disease Database (CHDD). The system designed in MATLAB software can be viewed as an alternative for existing methods to distinguish of heart disease presence.

Index terms: Heart disease, Data mining, Feature selection, classification, Modified Genetic Algorithm (MGA) Ensemble Online Sequential Extreme Learning Machine (EOS-ELM) , Cleveland Heart Disease Dataset (CHDD)

1. INTRODUCTION

"Data Mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data"[1]. In short, it is a process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different applications for example healthcare industry. Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from data. A major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly & provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable. According to survey of WHO, 17 million total global deaths are due to heart attacks and strokes. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. On the whole it is found as primary reason behind death in adults.

Among various life- threatening diseases, heart diseases have a great deal of attention in medical research. Also, it has more impact on human health. Various heart diseases was discussed and founded how they lead to heart attack [2]. The number one cause of death in industrialized countries was due to cardiovascular disease. Cardiovascular diseases not only have a major impact on individuals and their quality of life in general, but also on public health costs and the countries' economies. Risk factors for these pathologies include diabetes, smoking, family history, obesity, high cholesterol etc [3]. Health information decision was enabled by particularly knowing about the anatomy and functioning of the heart. A new born infant also have the possibility of heart disease. Some of the symptoms of heart disease in people were chest pain and fatigue. It occurred while the heart does not meet the circulatory demands of the body [4]. The physician takes decision based on the patient's answers to questions and lab results [5].

Blood flow to the heart muscles was decreased when block occurs in coronary arteries. The electrocardiogram recordings were analyzed to detect irregularity of heart beat problems occurred due to cardiovascular diseases [6]. In advance of medical and surgical treatment the patient with heart disease reached adulthood [7]. There are many diseases that affect the heart and arteries but four are particularly prevalent. Myocardial infarction was linked to damage to the coronary arteries in 90% of cases. Strokes

occurred as a result of impaired blood flow to the brain linked to a hemorrhage or a blockage of the arteries that supply blood to the brain. Heart failure was mainly linked to various changes in cardiovascular tissues, most often the result of ageing. High blood pressure was defined as the sustained elevation of arterial blood pressure in comparison to what is considered to be the "normal" value of 140/90 millimeters of mercury. There is a wide range of long-term consequences: heart failure, stroke, kidney failure etc.
The different types of heart disease widely in the world are Coronary heart disease, Heart failure, Coronary artery disease, Ischemic heart disease, Cardiovascular disease, Hypoplastic left heart syndrome, Atherosclerosis, Chronic obstructive pulmonary disease, Congenital heart disease, Valvular heart disease.

Mostly heart attacks are occurred when the plaque on the artery ruptures and a clot then forms, stopping blood flow. And the diagnosis of heart disease was based on medical knowledge occurred from patients. Correct diagnosis of the heart patient was delayed due to various problems. Diagnosis of heart disease was more costly and optimal decision path finder was used in terms of diagnostic accuracy while minimizing cost in diagnosis [8]. Heart disease can strike suddenly and quick decisions have to be made. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Prediction of heart diseases can provide some useful information about the health of patient. The prediction can be done with various computer aided diagnosis methods.

The datasets produced by different diagnostic procedures can be massive. The high dimensional nature of the data has given rise to a wealth of feature selection techniques being presented in this field. Feature selection is the method of selecting a feasible subset of features from the original set of candidate features. Unlike feature extraction, feature selection method is applied to datasets with known features. These methods will attempt to identify the significant features and discard irrelevant or redundant features from the original set of features. Feature selection methods can be classified into three major categories based on the technique of the search and selection process: complete, stochastic and heuristic search. Generally, artificial intelligence techniques were used in medical diagnosis with an improvement in prediction of heart disease [9]. Machine learning algorithm was used in medical diagnostic problem for heart disease. Case-based reasoning (CBR) was considered as a suitable technique for diagnosis, prognosis and prescription in the medical domain puts more stress on real cases than other domains. Coronary Artery Disease was diagnosed using two techniques called Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA) [9]. For the diagnosis of heart disease various classification and regression processes was used. It provides medical knowledge for diagnosis purpose [10]. So the dimensionality reduction is done by using Modified Genetic Algorithm (MGA) from these features set significantly speeds up the prediction task. This research paper added two more attributes i.e. obesity and smoking. The results from experiments returned with diminishing fact that there is considerable improvement in classification and prediction. The proposed EOS-ELM works as promising tool for prediction of heart disease

2. RELATED WORK

Jesmin et al [3] have proposed a computer intelligent based approach for the diagnosis of heart diseases. Apriori, Predictive Apriori and Tertius were the three different rule mining algorithms used to present rule extraction experiment on heart disease data and showed as efficiency algorithm for diagnosis task. Cleveland dataset, a publicly available dataset and widely popular with data mining researchers, have been used for diagnosis because of the privacy problem related to medical data set. Generally diagnosis were costly, time consuming and likely to suffer from error. The analyzed information available on sick and healthy individuals indicted that females have less chance of coronary heart disease than males. Heart disease for both men and women was existed only in the presence of exercise-induced angina and factors of men and women.

Jesmin et al [10] have examined the fact of computational intelligent techniques in heart disease diagnosis. Cleveland data was used to perform comparison with six well known classifiers. The potential of medical knowledge-driven feature selection was showed by comparing with computational intelligent based technique. And the imbalance data issue created by publicly available cleaved data was identified. For most classifiers and majority data set the performance was improved by the use of Motivated Feature Selection (MFS). It was because of the conversion of Cleveland data set for binary classification. The experimental results demonstrated that the use of MFS noticeably improved the performance, especially in terms of accuracy, for most of the classifiers considered and for majority of the datasets. MFS with Computer Feature Selection (CFS) was a promising technique used in heart disease diagnosis.

Tan et al [11] have proposed a hybrid approach consist of two conventional machine learning algorithms. Genetic Algorithms (GAs) and Support Vector Machines (SVMs) were the two proposed algorithm combined effectively based on a wrapper approach. Here, by an evolutionary process GA searches for the best attribute data set. Based on the attribute subset represented by GA, the SVM classified the patterns into reduced data set. This cyclic method was known as wrapper approach. UCI machine learning repository provided 5 set of data and it was checked by the proposed GA and SVM hybrid approach.

The GA-SVM hybrid approach attained an average accuracy of 76.20% which was relatively high. The robustness of the GA-SVM hybrid in the multi-class domain was showed by the obtained average accuracy 84.07%.

In the recent work developed a hybrid Neural Network [12] which included Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN). The proposed method accuracy, sensitivity and specificity measures were evaluated which were used commonly in medical classification. The aim of classification was to increase the reliability of the results obtained from the data. Here a new method was presented for classification of data of a medical database. The proposed method achieved accuracy values of 84.24% and 86.8% for Pima Indians diabetes dataset and Cleveland heart disease dataset respectively. The classification accuracies obtained by the proposed hybrid neural network were one of the best results compared with the results reported in the literature.

Particle Swarm Optimization (PSO)[13] is proposed for diagnosis of CHD. The proposed system uses Cleveland and Hungarian Heart Disease datasets. It includes four stages such as imputation of missing data, decision tree induction and rule extraction from imputed data set, using fuzzy membership functions, the crisp rules were transformed in to fuzzy rules and finally fuzzy membership functions were tuned by PSO. The generated Fuzzy Expert System (FES) based rules provides interpretation for the diagnosis of coronary heart disease. The approach has the ability to interpret the decisions made from the created FES. It provided 93.27% classification accuracy when compared to other approach

Cui et al [14] have proposed a training Artificial Neural Network (ANN) by exploiting Artificial Photosynthesis and Phototropism Mechanism (APPM). They used a stochastic optimization algorithm that stirs the plant growing process. In their algorithm each entity is called as branch and the sampled points are contemplated as branch growing trajectory. They have used two real world issues which are Cleveland heart disease categorization issue and sunspot number foreseeing issue to evaluate the performance of their APPM trained ANN. Their outcome showed that their technique increased the performance significantly contrast to other sophisticated machine learning techniques. Parthiban et al [15] projected an approach on basis of Coactive Neuro-Fuzzy Inference System (CANFIS) for prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease, were employed to carry out the experiment for the associative classifier.

3. PROPOSED METHODOLOGY

Most of the authors in data mining classifications techniques proposed for the prediction of heart disease, but the prediction system did not considered the uncertainty , unlabelled in the data measure and also multi class values. So, to remove the ambiguity, uncertainty, dimensionality of features, unlabelled and multiclass samples, we made an experiment with by of Online Sequential Extreme Learning Machine (EOS-ELM) classifier introducing a online functions to the classifier. The proposed EOS-ELM classifier results show promising in nature for removing the redundancy of data and to improve the accuracy of classifier as compared with other classifiers of supervised methods in data mining for Cleveland Heart Disease Database(CHDD) .

Patient Database : Patient database is datasets collected from Cleveland Heart Disease Dataset (CHDD) available on the UCI Repository [16]. The 15 attributes considered are age: age, sex, chest pain type, trestbps (resting blood pressure), chol (serum cholesterol in mg/dl), FBS (fasting blood sugar > 120 mg/dl), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), obesity ,smoking ,and CA (number of major vessels (0-3) colored by fluoroscopy). There are a total of 303 patient records in the database.

Data Preprocessing : This phase includes extraction of data from the Cleveland Heart Disease Dataset (CHDD) in a uniform format. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of outliers. Out of the 303 available records, 6 tuples have missing attributes. These have been excluded from the data set. For MCSSDB, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for decision trees or logistic regression.

Feature selection

But the selection of appropriate feature is challenging one. In this section the heart disease prediction system with evolutionary feature selection is explained. The main purpose of feature selection is to reduce the number of features used in clustering or classification while maintaining acceptable accuracy results for prediction. In this paper, Modified Genetic Algorithm (MGA) is proposed for dimensionality reduction .

Modified Genetic Algorithm (MGA)

Basic idea of GA is to imitate the mechanics of natural selection and genetics, one can make an analogy with the processes occurring in nature, saying that the probability that mutation takes place first and then comes crossover is comparable to the probability that both processes occur in a reverse order; or selection to be performed after crossover and mutation, no matter of their order. Following that idea, firstly implemented as a modified genetic algorithm SGA-CMS and applied to parameter identification of *E. coli* cultivation process [17], many modifications of SGA-SCM, concerning the sequence of execution of the main genetic operators, have been developed aiming to improve model accuracy and algorithm convergence time for the purposes of parameter identification of a fed-batch cultivation of *S. cerevisiae* [18-19]. SGA-CMS (crossover, mutation, selection), SGA-SMC (selection, mutation, crossover) and SGA-MCS (mutation, crossover, selection) have been proposed and thoroughly investigated in [20]. Two modifications skipping mutation operator – SGA-SC (selection, crossover) and SGA-CS (crossover, selection) have been also developed and applied [20].

Classification using EOS-ELM

In the study ambiguity, uncertainty, unlabelled samples for CHDD features and labelling measurement of CHDD features is determined by using Jensen-Shannon Divergence(JSD) .This work also supports multi-class prediction for heart diseases analysis with selected CHDD features. The MCSD model makes the heart disease prediction results for the pseudo labels for unlabeled CHDD features on the basis of both the prediction and the similarities among CHDD features which is determined from KLD. Considering $\mathrm{FS} = (\mathrm{chfs}_1, \ldots\ldots \mathrm{chfs}_N)$ to represent the set of CHDD features of N samples. Let us assume that the first $\mathrm{N_l}$ CHDD features are labeled by $\mathrm{y}_1, \ldots \mathrm{y}_{\mathrm{Nl}}$ where $\mathrm{y_i} = (\mathrm{y_i^1}, \ldots . \mathrm{y_i^m}) \in \{0, +1\}^{\mathrm{m}}$ is the binary vector and m is the number of classes for selected CHDD of dataset samples. $\mathrm{y_i^k} = +1$represents the selected CHDD features which is assigned to $\mathrm{k^{th}}$ class and $\mathrm{y_i^k} = 0$ is represented as the selected CHDD features which is not under the $\mathrm{k^{th}}$ class . The value $\hat{\mathrm{y}}_{\mathrm{i}} = (\hat{\mathrm{y}}_{\mathrm{i}}^1, \ldots \hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{m}}) \in \mathbb{R}^{\mathrm{m}}$ denotes the predicted class for selected CHDD features.

OS-ELM is developed on the basis of Extreme Learning Machine (ELM) [21] that is used for batch learning and has been shown to be extremely fast with good generalization performance. Compared to ELM,OS-ELM can learned at a one-by-one with fixed or varying chunk size. The parameters of hidden nodes in OS-ELM (input weights and biases for additive nodes or the centers and impact factors for RBF nodes) are randomly selected and the output weights are analytically determined. Simulation results in [21] have shown that OS-ELM is faster than other sequential algorithms and produces better generalization performances on many benchmark problems in the regression, classification and time-series prediction areas.

OS-ELM on the basis of ELM was developed for SLFNs with additive and RBF hidden nodes. Consider N arbitrary distinct samples $(x_i, t_i) \in R^n \times R^m$. If a SLFN with L hidden nodes can approximate these N samples with zero error ,it then implies that there exist b_i, a_i and b_i such that there exists β_i, a_i and b_i such that

$$f_L(x_j) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = t_j, j = 1, .. N \qquad (1)$$

where ai and bi are the learning parameters of the hidden nodes, bi is the output weight, and $G(a_i, b_i, x_j)$ denotes the output of the ith hidden node with respect to the input xj. When using additive hidden node, $G(a_i, b_i, x_j) = g(a_i . x_j + b_i), b_i \in R$, where a_i is the input weight vector, bi is the bias of the ith hidden node, and $a_i . x_j$ denotes the inner product of the two. When using RBF hidden node, $G(a_i, b_i, x_j) = g(b_i || x_j + b_i), b_i \in R^+$, where a_i and b_i are the center and impact width of the ith RBF node, and R^+, indicates the set of all positive real values.

Assume the network has L hidden nodes and the data. There are two phases in OS-ELM algorithm ,an initialization phase and a sequential phase. In the initialization phase, rank H0= L is required to ensure that OS-ELM can achieve the same learning performance as ELM, where H0 denotes the hidden output matrix for initialization phase .It means the number of training data required in the initialization phase N0 has to be equal to or greater than L, i.e. $N_0 \geq L$. And if N0 = N, OS-ELM is the same as batch ELM. Hence, ELM can be seen as a special case of OS-ELM when all the data present in one iteration.

(a) Randomly assign the input parameters: for additive hidden nodes, parameters are input weights ai and bias bi; for RBF hidden nodes ,parameters are center ai and impact factor bi; $i = (1, \ldots L)$.

(b) Calculating the initial hidden layer output matrix H0

$$H_0 = \begin{bmatrix} G(a_1, b_1, x_1) & \ldots & G(a_L, b_L, x_L) \\ \vdots & & \\ G(a_1, b_1, x_{N_0}) & & G(a_L, b_L, x_{N_0}) \end{bmatrix}_{N_0 \times L} \qquad (2)$$

Estimating the initial output weight β^0. Set k = 0. (k: a parameter indicates the number of chunks of data that is presented to the network.)

Sequential learning phase :Present the (k+1) th chunk of new observations

$$N_{k+1} = \{(x_i, t_i)\}_{i=(\sum_{j=0}^{k} N_j)+1}^{\sum_{j=0}^{k+1} N_j} \qquad (3)$$

and N_{k+1} denotes the number of observation in the (k+1) the chunk .Compute the partial hidden layer output matrix H_{k+1}

$$H_{k+1} = \begin{bmatrix} G\left(a_1, b_1, x_{\left(\sum_{j=0}^{k} N_j\right)+1}\right) & \cdots & G\left(a_L, b_L, x_{\left(\sum_{j=0}^{k} N_j\right)+1,}\right. \\ \vdots & & \\ G\left(a_1, b_1, x_{\left(\sum_{j=0}^{k} N_j\right)+1}\right) & & G\left(a_L, b_L, x_{\left(\sum_{j=0}^{k} N_j\right)+1,}\right. \end{bmatrix} \quad (4)$$

Calculate the output weight $\beta^{(k+1)}$ .Have $T_{k+1} = \left[t_{(\sum_{j=0}^{k+1} N_j)\ +1,\dots,} t_{(\sum_{j=0}^{k+1} N_j)}\right]^T_{N_{k+1 \times m}}$

Set k=k+1 , go to (a) in this sequential learning phase.

Ensemble of OS-ELM EOS-ELM consists of many OS-ELM networks with same number of hidden nodes and same activation function for each hidden node. Constructed P OS-ELM networks to form our EOS-ELM. All P OS-ELMs are trained with new data in each incremental step. The input parameters for each OS-ELM network are randomly generated and the output weights are obtained analytically based on the sequential arrived input data. Then compute the average of the outputs of each OS-ELM network, which is the final output of the EOS-ELM. Assume the output of each OS-ELM network is $f^{(i)}(x_i), j = 1,, . P$. Hence,

$$f(x_i) = \frac{1}{P}\sum_{j=1}^{P} f^{(j)}(x_i) \quad (5)$$

Expect that EOS-ELM works better than individual OS-ELM network because the randomly generated parameters make each OS-ELM network in the ensemble distinct. Therefore, the OS-ELM networks composing the ensemble may have different adaptive capacity to the new data. When the data come into the ensemble network sequentially, some of OS-ELM networks may adapt faster and better to the new data than others. However, because for different incoming data, different OS-ELM networks can be the ones that have good adaptation. When $N_0 = N$, EOS-ELM becomes an ensemble of batch ELM networks [22]. Therefore, the ensemble of ELM proposed in [22] can be seen as a special case of EOS-ELM when all the training data are available at one time

4. EXPERIMENTATION RESULTS

The data set is taken from the Data Mining Repository of the University of California, Irvine (UCI) [16]. To end with the system is tested using Cleveland data sets. Attributes such as Age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved, exercise induced angina, ST depression, and slope of the peak exercise ST segment, number of major vessels, thal and the diagnosis of heart disease are presented. In experimentation work we have used a total of 909 records with 15 medical attributes. This dataset is taken from Cleveland Heart Disease database [16].Have split this record into two categories: one is training dataset (455 records) and second is testing dataset (454 records). The records for each category are selected randomly. “Diagnosis” attribute is the target predictable attribute. Value “1” of this attribute for patients with heart disease and value “0” for patients with no heart disease. “PatientID” is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved. Table 1 shows the attribute information of Cleveland Heart Disease database. There are multi classes to be predicted: Absence, Moderate and presence of heart disease in patients.

Table 1: Attribute information of Cleveland Heart Disease database

Attributes	Description	Type
Age	age in years	Numerical
sex	sex (1 = male; 0 = female)	Categorical
cp	chest pain type • Value 1: typical angina • Value 2: atypical angina • Value 3: non-anginal pain • Value 4: asymptomatic	Categorical
restbps	resting blood pressure (in mm Hg on admission to the hospital	Numerical
chol	serum cholestoral in mg/dl #10 (trestbps)	Numerical
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Categorical
restecg:	resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	maximum heart rate achieved	Numerical
exang	exercise induced angina (1 = yes; 0 = no)	Categorical
Oldpeak	ST depression induced by	Numerical

	exercise relative to rest	
slope	the slope of the peak exercise ST segment Value 1: upsloping, Value 2: flat and Value 3: downsloping	Categorical
ca:	number of major vessels (0-3) colored by flourosopy	Categorical
thal	3 = normal; 6 = fixed defect; 7 = reversable defect	Categorical
num:	diagnosis of heart disease Value 1: present Value 0: not_present	Categorical
obes	1 = yes 0 = no	Categorical
smoke	1= past 2 = current 3 = never	Categorical

Accuracy is the typically used measure to evaluate the efficacy of clustering methods; it is used to reckon how the test was worthy and consistent. In order to calculate these metric, we first compute some of the terms like, True positive (TP), True negative (TN), False negative (FN) and False positive (FP) based on Table 2.

Table 2.Confusion matrix

Result of the diagnostic test		Physician diagnosis	
		Positive	Negative
Clustering results	Positive	TP	FP
	Negative	FN	TN

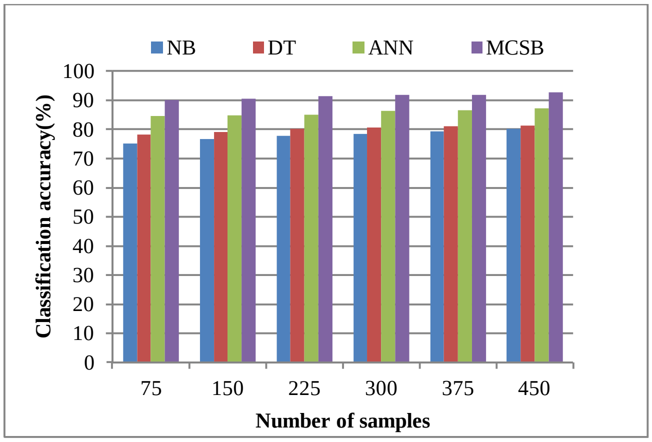

Figure 1: Classification accuracy comparison for methods vs number of samples

From Figure 1 shows the prediction accuracy results of various classification methods for Cleveland database with 15 attributes. It shows that the prediction accuracy results of the proposed EOS-ELM schema is increases when compare to existing classification methods since the proposed work missing attribute data is replaced with the help of preprocessing methods and feature selection is done using FSE method which reduces error rate of the classifier.

The prediction accuracy results are also measured and evaluated using the following metrics such as PE, V-Measure [23].

Partition Entropy Coefficient (PE): Many unsupervised evaluation measures have been defined, but most are only applicable to clusters represented using prototypes. Two exceptions are the Partition Coefficient (PC) and the closely related Partition Entropy Coefficient (6) , the latter of which is defined as,

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|L|}(u_{ij}\log_a u_{ij}) \quad (6)$$

where $u_{ij}$is the membership of instance i to cluster j. The value of this index ranges from 0 to $\log_a|L|$. The closer the value is to 0, the crisper the clustering is. The highest value is obtained when all of the $u_{ij}$is are equal. The remainder of the criteria that we describe is all supervised.

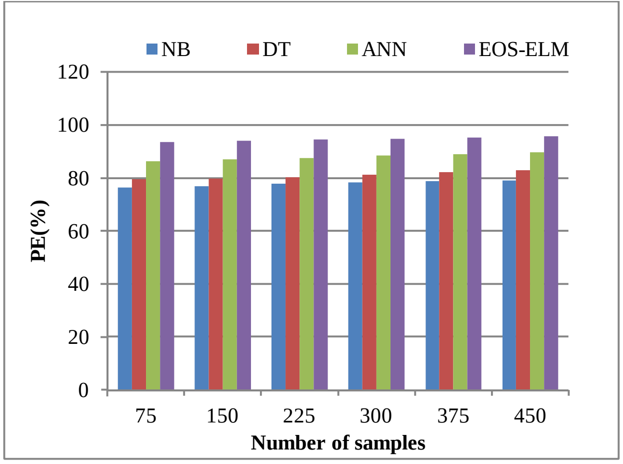

Figure 2: PE comparison for methods vs number of samples

From Figure 2 shows the PE prediction accuracy results of various classification methods for Cleveland database with 15 attributes. It shows that the PE results of the proposed EOS-ELM schema is increases when compare to existing classification methods since the proposed work missing attribute data is replaced with the help of pre-processing methods and feature selection is done using FSE which reduces error rate of the classifier.

V-Measure [23]: This problem with purity and entropy is overcome by the V –measure (9) , also known as the Normalized Mutual Information (NMI) , which is defined as the harmonic mean of homogeneity (h) and completeness (c); i.e.,

$$V = \frac{hc}{h + c} \quad (7)$$

where h (12) and c are defined as

$$h = 1 - \frac{H(C|L)}{H(C)} \; \& \; c = 1 - \frac{H(L|C)}{H(L)} \qquad (8)$$

$$H(C) = -\sum_{i=1}^{|C|} \frac{|c_i|}{N} \log \frac{|c_i|}{N} \qquad (9)$$

$$H(L) = -\sum_{i=1}^{|L|} \frac{|w_i|}{N} \log \frac{|w_j|}{N} \qquad (10)$$

$$H(C|L) = -\sum_{j=1}^{|L|} \sum_{i=1}^{|C|} \frac{|W_j \cap c_i|}{N} \log \frac{|W_j \cap c_i|}{|c_i|} \qquad (11)$$

$$H(L|C) = -\sum_{i=1}^{|C|} \sum_{j=1}^{|L|} \frac{|w_j \cap c_i|}{N} \log \frac{|w_j \cap c_i|}{|c_i|} \qquad (12)$$

Because it takes into account both homogeneity and completeness, V -measure is more reliable than purity or entropy when comparing clusterings with different numbers of clusters.

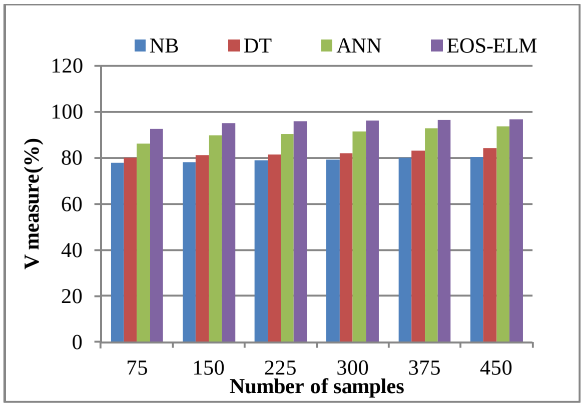

Figure 3: V-Measure comparison for methods vs number of samples

From Figure 3 shows the V-Measure results of various classification methods for Cleveland database with 15 attributes. It shows that the V-Measure results of the proposed **EOS-ELM** schema is increases when compare to existing classification methods since the proposed work missing attribute data is replaced with the help of pre-processing methods and feature selection is done using FSE which reduces error rate of the classifier.

5. CONCLUSION AND FUTURE WORK

The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. The diagnosis of heart disease is a significant and tedious task in medicine. This research paper proposed a feature selection method for Heart Disease Prediction. This research paper added two more attributes i.e. obesity and smoking. The results from experiments returned with diminishing fact that there is considerable improvement in classification and prediction. The proposed EOS-ELM works as promising tool for prediction of heart disease when compared to other data mining classification techniques, Also from the accuracy perspectives, seven techniques provide more than 95% accuracy as compared with the other techniques presented in the literature. Finally, some of the research issue is also addressed to preceed the further research in the same direction.

REFERENCES

1. Frawley and G. Piatetsky -Shapiro, Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.
2. Jae-Hong Eom and Sung-Chun Kim, et al, "AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction", Journal of Expert Systems with Applications, Vol. 34 2465, PP.2479, 2008
3. Jesmin Nahar and Tasadduq Imam et al," Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013
4. Chang-Sik Son and Yoon-Nyun Kim, et al, "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, Vol.45, PP. 999–1008,2012
5. Hongmei Yan and Jun Zheng, et al, "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm", Journal of Applied Soft Computing, Vol.8, PP.1105-1111, 2008
6. Swati Shilaskar et al, Feature selection for medical diagnosis: Evaluation for cardiovascular diseases", Journal of Expert System with Application, Vol.40, PP.4146-4153, 2013
7. Petra A. Karsdorp and Merel Kindt et al, "False Heart Rate Feedback and the Perception of Heart Symptoms in Patients with Congenital Heart Disease and Anxiety", International Journal of behavioral Medicine, Vol.16, PP.81-88, 2009
8. Chih-Lin Chi and W. Nick Street, et al, "A decision support system for cost-effective diagnosis", Journal of Artificial Intelligence in Medicine, Vol.50, PP. 149-161, 2010.
9. Ismail Babaoglu and Og˘uz Findik et al, "A comparison of feature selection models utilizing

binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine", Journal of Expert System With Applications, Vol.37, PP.3177-3183, 2010

10. Jesmin Nahar and Tasadduq Imam, et al, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach", Journal of Expert System with Application, Vol.40, PP.96-104, 2013
11. K.C. Tan and E.J. Teoh et al, "A hybrid evolutionary algorithm for attribute selection in data mining", Journal of Expert system with applications, Vol.36, PP.8616-8630, 2009
12. Humar Kahramanli and Novruz Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", Journal of Expert Systems with Applications, Vol. 35, PP. 82–89, 2008
13. S. Muthukaruppan and M.J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease", Journal of Expert Systems with Applications, Vol.39, PP. 11657–11665, 2012
14. Zhihua Cui et al, "Training artificial neural networks using APPM", International Journal of wireless and mobile computing, Vol.5, PP.168-174, 2012.
15. Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).
16. Anamika Gupta, Naveen Kumar, and VasudhaBhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings Of World Academy Of Science, Engineering And Technology,Vol. 6, June 2005 .
17. O. Roeva, A Modified Genetic Algorithm for a Parameter Identification of Fermentation Processes, Biotechnology and Biotechnological Equipment, Vol.20, No.1, 2006, pp. 202-209
18. M. Angelova, T. Pencheva, Algorithms Improving Convergence Time in Parameter Identification of Fed-Batch Cultivation, Comptes rendus de l'Academie bulgare des Sciences, 2012, Vol.65, No.3, pp. 299-306.
19. M. Angelova, T. Pencheva, Tuning Genetic Algorithm Parameters to Improve Convergence Time, International Journal of Chemical Engineering, 2011, Article ID 646917, available at http:// www.hindawi.com /journals/ ijce/2011/646917/
20. M. Angelova, S. Tzonkov, T. Pencheva, Genetic Algorithms based Parameter Identification of Yeast Fed-Batch Cultivation, Proceedings of the Conference on "Numerical Methods and Applications", LNCS, Vol.6046, 2011, pp. 224-231.
21. G.-B.Huang, L.Chen, C.-K.Siew, Universal approximation using incremental constructive feed forward networks with random hidden nodes ,IEEE Transactions, Neural Networks, 17(4)(2006)879–892.
22. G.H.Golub , C.F.V.Loan, Matrix Computations ,third ed., The Johns Hopkins University Press, Baltimore, MD,1996.
23. Z.-L.Sun,T.-M.Choi,K.-F.Au, Y.Yu, Sales forecasting using extreme learning machine with applications in fashion retailing, Decision Support Systems 46 (1) (2008)411–419
24. Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In EMNLP-CoNLL (Vol. 7, pp. 410-420).