South Asian Journal of Engineering and Technology Vol.2, No.15 (2016) 51–57

ISSN No: 2454-9614

Classification of Medical Dataset using Soft computing Techniques

J. Preethi[a], K. DeviPriya[a*], S. Dhivya[a,b]

[a)] *Department of Computer Science and Engineering, Anna University Regional Campus Coimbatore,Tamilnadu, India.*
[b)] *Department of Computer Science and Engineering, Anna University Regional Campus Coimbatore, Tamilnadu, India.*

*Corresponding Author: J. Preethi

E-mail:preethi17j@yahoo.com,
Received: 03/11/2015, Revised: 20/12/2015 and Accepted: 27/02/2016

Abstract

In this paper, the application of Artificial intelligence computing techniques for diagnostic of disease by classifying the biomedical datasets. Many Artificial Intelligence techniques were reviewed for medical dataset classification. This Exploration assembles typical work that shows how the Artificial Neural Network is applied to the solution of different diagnostic disease with classification. It also detects the methods and the techniques of ANN that are used frequently to solve the special problem related to the medical dataset classification. Extreme Learning Machine (ELM) is used in almost for learning the medical datasets to the network. Similarly PSO is mainly used for datasets classification of parameters or attributes. Several diseases like Breast cancer, Heart disease, Diabetes….etc using ANN approach are result in a use of SVM (Support vector Machine) and BP network.

Reviewed by* **ICETSET'16 *organizing committee*
Keywords: Medical datasets, Machine learning, Artificial Intelligence (AI), Extreme learning Machine (ELM), Artificial Neural Network (ANN)

1. Introduction

Artificial Intelligence (AI) - It is the science of making intelligent machines, especially intelligent programs. And diagnosis is followed with the development of algorithm and techniques that are able to determine whether the behavior of a system is correct. The application of computational or machine intelligence in medical diagnosis is a new trend for medical dataset classification. Classification system mainly helps to minimize possible errors that can be done because of inexperienced technicians, and also provide medical datasets which can be examined in the shorter time and in detail.

Automated diagnostic systems are one of the areas of analyzing database and medical dataset classification. The objective of these studies is assisting the doctors in making a diagnostic decision with a subject to assure the diagnosis aid precisely. That the classification efficiency is allowed through Sensitivity, Specificity, and Accuracy for classification functions. A good classifier should give hundred percent results for all the three.

2. Medical Dataset

Medical classification, or medical coding, is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. The diagnoses and procedures within the healthcare record, such as the transcription of the physician's notes, laboratory results, radiologic results, and other sources are usually taken from a variety of source. That the medical datasets are taken from the UCI machine repository.

LIST OF DISEASES	NUMBER OF INSTANCES	NUMBER OF ATTRIBUTES	NUMBER OF CLASSES
BREAST CANCER	699	10	2
DIABETES	1151	20	3
LUNG CANCER	32	56	4
HEART	303	75	2
HEPATITIS	155	19	2
THYROID	7200	21	4

Table 1 - Different medical datasets

3. Overall Structure

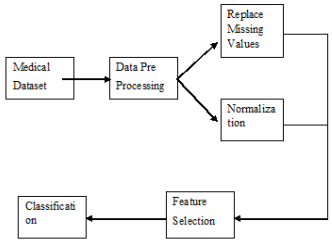

Figure 1: Overall Structure

4. Literature Review

4.1. Medical Dataset Analysis Methods

Medical dataset analysis method is included with three important steps. First step includes dataset pre-processing, second follows feature selection and final step include classifying the dataset. This all described in following sections

4.2. Dataset Pre-processing

Pre-processing the thousands of medical datasets are combined into one relational table, by pre-processing it delete the mismatched data and also it removes the multivalve attribute. And it replaces the missing value by its mean, median and its standard deviation (SD). Normalization is used to bring the datasets within a range (e.g. from -1 to 1) and Min-Max normalization is used by,

$$x_i^j = \frac{x_i^j - min^j}{max^j - min^j} \quad (1)$$

Where, x_i^j is the initial value and max^j ,min^j is the mean and standard deviation.

4.3. Feature Selection

After normalizing the datasets, Feature selection method is used to get the most important features of the dataset. Several methods are used for feature selection are F-Score, Threshold fuzzy entropy, PCA, GDA etc.

4.3.1. F-Score

Feature selection by F-Score is used to find the optimal subset of input variable with the best feature by removing no predictive information. That it gives the good accuracy value for classifying the medical datasets. It follows some steps for feature selection.

- All features are taken and calculated using the given formula.
- Mean is calculated using the formula.
- Compare the mean value of the feature with the original value.
- It measures the discrimination or relevant feature values related to 2 different features.

4.3.1.1 Related Works

This paper deals with the F-Score feature selection method with Support Vector Machine classifier. That F-Score is calculated for each and every feature using the given F-Score formula, And it calculate the discrimination between the two real numbers with positive and negative numbers for all given features,

$$f_i = \frac{((x_i^{(+)} - x_i) + (x_i^{(-)} - x_i))}{n_+^{-1} \sum_{k=1}^{n} \left(x_k^{(+)} - x_i^-\right) + \frac{i}{n-1}\left(x_k^{(+)} - x_i^-\right)}$$

Where, the numerator deals with the average of the ith value, And the denominator deals with the discrimination between two different points of the sets. *Peng Tao et al.,[2]* Weighted F-Score method by reducing the dimensional space of the feature is selected. It calculates the interclass coefficient by the given formula which is based on traditional F-Score method. The given formula tells the number of points in k^{th} and q^{th} features of the given input samples,

The feature selection is demonstrated by,

METHOD	LIST OF DATASETS	ORIGINAL FEATURE	REDUCED FEATURE
IMPROVED F-SCORE	BREAST CANCER	9	4
WEIGHTED F-SCORE	HEPATITIS	12	6
F-SCORE	DIABETICS	10	8

Table 2: Different approach for Feature Selection

4.3.2. Threshold Fuzzy Entropy based Feature Selection

Feature is selected based on the Fuzzy C-means Clustering with three frameworks is followed

- Mean selection
- Half selection
- Neural network

4.3.2.1. Related Works

P. Jaganathan el al.,[16] this paper deals with fuzzy entropy which same as the FCM (Fuzzy C Means) , thus it uses the membership function to each and every input features. By using the membership function the clusters are grouped. That similar features are grouped into one cluster and dissimilar feature are grouped into one cluster. The dissimilar features are identified using the Elucdiean distance. For feature selection FCM algorithm is used as follow,

- It assumes the cluster size and data points with the condition 2<c<N, where N is the number of clusters.
- It calculate the membership function by,

$$C_j = \frac{\sum_{i=1}^{N} u_{ij} x_{ij}}{\sum_{i=1}^{N} u_{ij} g_{ij}} \quad (3)$$

- Elucdiean distance is calculated by,

$$D_{ij} = c_j - x_i \quad (4)$$

- Update fuzzy values according to the calculated distance

4.3.3. PCA

PCA used to convert the set of observation of possibly correlated variable into the set of linearly uncorrelated values. And it reduces the dimension of the datasets with minimal loss of information and selects the most relevant feature. The reduction of feature dimension by extract the sub set of feature that describe as the best feature and evaluated with high accuracy.

4.3.3.1. Related Works

*Kemal Polat et al., **[17]*** this paper deals with principal component analysis that gives more accuracy by bring out a subset of each features of the diabetics datasets. Features of the diabetic datasets are allocated with sample covariance matrix,

$$ST_i = \gamma_i T_i \quad (5)$$

The principal component analysis *m*, with largest Eigen vector λi value of xi. And m PC of x are decorrelated in input space and mapped to the global covariance matrix.

$$ST_i = \left(\frac{i}{N}\right) \sum_{i=1}^{N} (x_i - \mu) + (x_i - \mu) \quad (6)$$

U global mean of the matrix, K is number of classes taken for diabetics dataset, Nj is the number of samples taken.

By performing the PCA the classification accuracy of diabetic reaches 89.87% with 10 fold cross validation test.

4.3.4. GDA

General Discriminate Analysis is used as a linear analysis model to the discriminate analysis problem and also for classification problem. By using GDA we can set a complex model for the set of predictor variable.

4.3.4.1. Related Work

Salih Gunes et al., [18] this paper deals with non-linear classification with the given vector matrix which is based on the kernel function O. And the kernel function transforms or converts the original vector space to reduced new dimensional space.

Z : U : X! Z (7)

5. Feature Classification

After the feature is extracted using the feature selection techniques. The best feature is selected using the classification algorithms like Neural Network, Support Vector Machine (SVM), K-NN (K-Nearest Neighborhood) , Decision Tree. By using the following techniques the best feature are selected and grouped into one and classifying the datasets with high accurate of disease rate. That classification having both the supervised and unsupervised learning algorithms. Classification techniques for supervised learning are Cluster, Fuzzy C Means etc... And unsupervised learning algorithms are ANN, K-NN, Decision tree etc…

5.1 Neural Network

Artificial Neural Network is a powerful tool for solving the complex problem with linear input-output efficient relationship. That neural network is based on connections with human brain which having N number of neurons. Neural network are followed with three layers they are input layer, output layer and hidden layer. Hidden layer is used to map the input function to the output. Neural network having following features it support several network architecture for supervised and unsupervised learning, it uses parallel computing for feature training process, dynamic network to store the data.

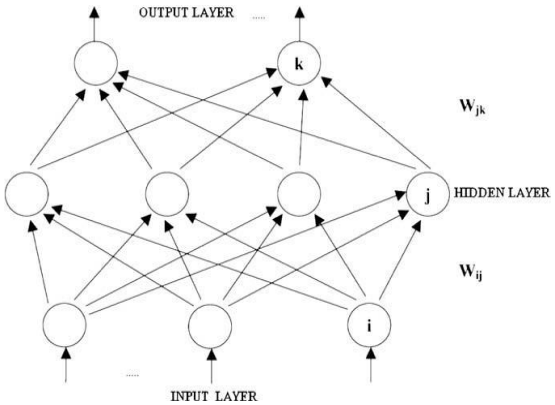

Figure 3: Neural Network

5.1.1. Related works

You Xu et al., **[1]** this paper deals with the Single hidden layer feed forward network with off learning called ELM is used. That SLFF with ELM learn the hidden parameter and input weights randomly while the output layer is calculated using MP inverse method. *Guang-Bin Huang et al*.**,[4]** this paper deals with Extreme Learning machine with single layer feed forward network, by using the SLFN two functions are followed (i) it approximate the complex mapping from the input space. (ii) And it provide the complex model for classical parameter.

5.2. SVM

Classification with SVM by finding the hyper plane that increases the dimensions of two plane classes. The hyper plane vector are defines support vectors.

Maximize the width of the margin to get the optimal hyper plane. SVM are allowed with linear hyper plane, and then it has a unique global value. It built a model of new sample into one class, making a non probabilistic linear classifier.

5.2.1. Related works

Mehmet Fatih Akay et al., **[13]** this paper deals with how the SVM are allowed as a classifier for the medical datasets. That it is used to calculate the boundary between two classes.

Salih Gunes et al., **[18]** it deals the SVM based on statistical learning theory which is used to train the network and it is used for classification and regression. And the classification with SVM is done using the support vectors. It achieves the classification only by linear case and not with non linear case.

5.3. K-NN

K-NN is the non linear parameter used for classification and regression. In both, for classification and regression the input are combined with K closest neighbor values of the feature space. Classification is followed according to the classification condition. The output is assigned by membership class. That the value of K is either Positive or small. They classify the object or feature according to the value.

5.3.1. Related work

Seral , Sahanet al.**, [12]** this paper deals with the K-NN algorithm for classification That K-NN is the instance based learning algorithm for classification problem. The classifying unit consist of samples that are related to the system. It follows the n dimensional space assumed to the entire instance. That with dimensional space the distance between the points or the instances is measured using the Euclidean distance with That all the samples are followed with f(.) function, it allows all the sample in the system unit. Related classes of the training sets are also a system unit.

6. Results

The Diagnosis of the diseases is identified by medical datasets with several feature selection methods and classification techniques. The diagnosis accuracy may be varied from one exact technique to another technique. Different feature extraction techniques that combine with the classification method provide better results.

The Table 3 shows that the various accuracy levels of feature extraction and classification methods during the emotion identification process.

S NO	SELECTION METHOD	CLASSIFICATION ALGORITHM	ACCURACY	CITIATION
1	Kernel F-Score	SVM	76.03%	[1]
2	Fuzzy entropy	K-NN	75.45%	[7]
3	F –Score	NN	85.90%	[3]
4	Weighted F-Score	RBFsss	79.12%	[8]
5	PCA	SVM	67.8%	[3,4]
6	GDA	NN	70.8%	[5]

Table 3: Accuracy of Methods

7. Conclusion

Classification of medical dataset can be identified by extracting the different kind of feature from the datasets. For extracting the features from the datasets Fuzzy C Means Clustering will give the highest accuracy, after pre-processing the signal it has to be smoothed and optimized for the particular feature, using different optimization techniques like SVM, PSO, etc.. After getting the optimized result apply the PSO and Fuzzy Cognitive map it will provide the high accuracy of classification. These classification of medical dataset will used for diagnosis of disease in early stage.

References

[1] Kemal Polat *, Salih Gunes ”A new feature selection method on classification of medical datasets: Kernel F-score feature selection” in 2009 Elsevier.

[2] P. Jaganathan and R. Kuppuchamy, “A threshold fuzzy entropy based feature selection for medical database classification,”*Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2222– 2229, 2013.

[3] Peng Tao1, Huang Yi” A Method Based on Weighted F-score and SVM for Feature Selection” in 2015 CCDC.

[4] H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks, *Expert Systems with Applications*, vol. 36, no. 4, pp. 8610–8615, 2009.

[5] K. Polat, S.G¨unes¸, andA.Arslan, “Acascade learning systemfor classification of diabetes disease: generalized discriminant analysis and least square support vector machine,” *Expert Systemswith Applications*, vol. 34, no. 1, pp. 482–487