ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Novel Machine Learning Approach to Diagnose Type 2 Diabetes and Different Clinical Datasets Using Weighted Genetic PCA Methods

B.Senthil Kumar[1], Sreejith.R[2]

Assistant Professor, Dept. of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[1]

M.Phil Scholar, Dept. of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[2].

ABSTRACT: The Intelligent health care systems are performing the data mining techniques for effective data analysis and management also. Diagnosis and detection of diseases from patient electronic health records are very dynamic in nature and achieving that is a very promising area of research. From the numerous health data, the proposed system handles two popular disease dataset such as Diabetes and heart diseases. The proposed system diagnoses diabetes and heart diseases with its risk factors from dynamic huge volume health records. Even though there are several approaches in data mining is introduced, still some issues arises while classifying. So, the system proposes a new iterative approach, which concentrates on the effective feature selection and classification. The classification accuracy depends on the data which given at the time of training. The existing system suffers from improper training dataset, so it created a class imbalance problem. To handle such issues, the system concentrates on three main portions for accurate classification. One is Effective pre-processing, feature selection and classification. The pre-processing stage eliminates the inconsistent and redundant dataset and reduces the class imbalance problems. The second stage is the feature selection process, which performed using PCA (Principle Component Analysis) and effective classification using WGA (Weighted Genetic Algorithm). The system implements a new weighted Genetic based algorithm with the use of effective weighted features from the PCA. The system finds the type2 Diabetes and Heart disease Classification using WGA technique. The system developed with the intension of high accuracy and less training overhead.

KEYWORDS: Type 2 Diabetes, Heart Disease, PCA, Genetic algorithm, Machine Learning, data mining, classification and prediction.

I. INTRODUCTION

Data mining is an integration of multiple disciplines such as statistics, machine learning, neural networks and pattern recognition. It is concerned with the process of computationally extracting hidden knowledge structures represented in models and patterns from large data repositories. Healthcare is a data intensive process. Many processes run simultaneously producing new data every second. It is a research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This has resulted in unification of healthcare and computing to form Health care. They typically work through an analysis of medical data and a knowledge base of clinical expertise and it is an emerging field. In [1] authors described the need and algorithms of data mining in healthcare, in medical areas today, data collection about different diseases as very important. Medical and health areas are among the most important sections in industrial societies. The extraction of knowledge from a massive volume of data related to diseases and medical records using the data mining process can lead to identifying the laws governing the creation, the development of epidemic diseases [2][3].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

II. PROBLEM DEFINITION

Discovery of new information in terms of patterns or rules from large amounts of data is based on the machine learning technique [4]. Disease prediction plays an important role in data mining. Diagnosis of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can reduce the number of tests. This reduced test set plays an important role in time and performance. Diabetes data mining is important because it allows doctors to see which features or attributes are more important for diagnosis such as age, weight, etc. This will help the doctors diagnose diabetes more efficiently. There are various data mining techniques in use in healthcare industry but the research that has to be done is on the performance of the various classification techniques, to enable the choice of the best among them can be chosen.

The research presented in this thesis is intended to address the challenge of improving the prediction model to predict the heart disease and diabetic disease and providing timely response in predicting the disease. Briefly the important research functions are therefore stated as,

• Various datasets are used in the proposed classifier and prediction technique.

• A classification techniques help in developing the prediction model so as to predict accurately the risk of heart disease among diabetic patients.

III. PROPOSED SYSTEM

Application of data mining in analyzing the medical data is a good method for investigating the existing relationships between variables. Nowadays, data stored in medical databases are growing in an increasingly rapid rate. It has been widely recognized that medical data analysis can lead to an enhancement of health care. The primary objective of the research work is the effective development of prediction model using various classification techniques to predict the diabetes and heart disease and performance in prediction. It also shows that data mining can be applied to the medical databases to predict or classify the data with reasonable accuracy. The following are the objectives leading to achievement of the primary objective mentioned supra:

• To generate a best classification technique which can help in predicting the risk of heart and diabetic disease with various attributes.

• To recognize and classify patterns in multivariate patient attributes.

• To predict the class score based on that, the mild and extreme of the disease can be identified.

• To improve the classification and prediction accuracy by utilizing improved classification techniques.

To propose a weighted Genetic Algorithm and PCA (Principle Component Analysis) [5] based feature selection approach and compare the performance of existing and the proposed feature selection algorithms on clinical datasets.

To design a fusion based Classifier for diabetes and heart disease diagnosis and to predict the severity of heart disease in patients. To propose a scoring system to find the severity of heart disease for patients who are suffering with diabetes.

The system proposes a new iterative approach, which concentrates on the effective feature selection using PCA (Principle Component Analysis) and effective classification using WGA (Weighted Genetic Algorithm). This chapter discuss about the algorithms and methodologies. The followings are the main contributions of the proposed work.

- In this paper, we implemented a new Genetic based algorithm with the use of effective weighted features from the PCA. The system introduces a new Diabetes and Heart disease Classification algorithm with WGA technique.
- We also created a new advanced classification for fast disease classification. The system developed with the intension of high accuracy and less training overhead.

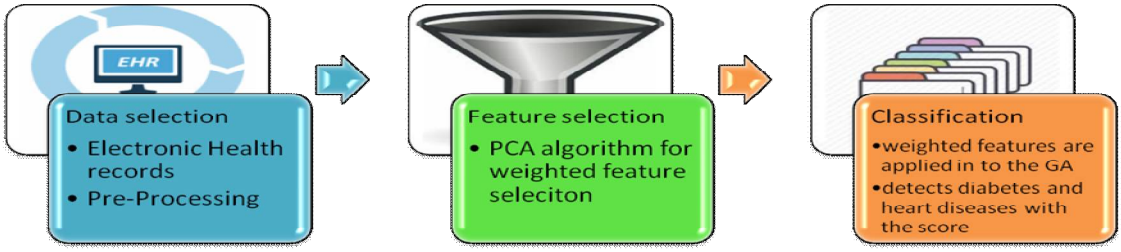

Fig 1.0 the overall process involved with the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

- Diabetes and Heart disease classification and prediction of the class score using a renovation algorithm which is a combination of PCA (Weighted Principle Component Analysis) and weighted genetic algorithms.
- PCA for feature selection and dimensionality reduction
- And Genetic algorithm for disease classification and prediction.

The optimized PCA algorithm has been expanded with the new optimal classification algorithms, which can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time.

A. DATA SELECTION:

The real world data is incomplete, noisy and inconsistent. Data pre-processing routines attends to filling the missing values, smooth the noisy data while identifying outliers and correct inconsistency in data. Noise is a random error or variance in a measured variable. Binning can be used for noise removal. Binning methods smooth a sorted data value by consulting its neighborhood. The sorted values are distributed into a number of buckets or bins. At the time of data selection, each value of the attribute is replaced by the mean value of the selected attribute. And this also includes the minimum and maximum values of each attribute values and this will be used as boundaries. Each value is then replaced by the closest value for every attribute.

B. FEATURE SELECTION USING PRINCIPLE COMPONENT ANALYSIS:

Feature selection is a process commonly used in the second stage of the proposed work, wherein a subset of the features available from the data is selected for application and that will be applied into the learning algorithm. The finest and suitable subset will be selected, that should contain the least number of dimensions. If the dimensions are less the accuracy of the classifier will be high. This discards the remaining, unimportant dimensions using PCA. This is an important stage after pre-processing and is one of two ways of avoiding the curse of dimensionality in the health care domain.

There are two approaches in feature selection namely forward selection and backward selection in the enhanced PCA. Forward Selection starts with no variables and adds them one by one at every iteration, this will reduces the error. Backward Selection starts with all the variables and removes them one by one. In the feature selection process, all subset selection evaluates the features as a group for suitability. Subset selection algorithms can be segmented into Wrappers, Filters and Embedded. In this research work, PCA, Genetic Algorithm, have been implemented which are explained in detail in this section.

Algorithm: Feature selection Algorithm:

Input: patient dataset
Output: Diagnosed result with risk level and performance result
Steps:

1. Taking the whole dataset ignoring the class labels
2. Find initial component
3. Compute the *d*-dimensional mean vector
4. Compute the covariance matrix of the original or standardized d-dimensional dataset X (here: d=3); alternatively, compute the correlation matrix.
5. Eigen-decomposition: Compute the eigenvectors and eigenvalues of the covariance matrix (or correlation matrix).
6. Sort the eigenvalues in descending order.
7. Choose the k eigenvectors that match to the k^{th} largest Eigen values, here k is the number of features of the new feature subspace ($k \leq d$).
8. Construct the projection matrix W from the k selected eigenvectors.
9. Transform the original dataset X to obtain the k dimensional feature subspace Y ($Y=WT \cdot X$).
10. Return features for classification

Algorithm1: Feature selection steps

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

C. CLASSIFICATION AND CLASS RANGE PREDICTION FOR HEALTH DATASETS:

A genetic algorithm is a type of searching algorithm and overcomes the optimization problem. GA searches a solution space for an optimal solution to a problem using three steps. Genetic Algorithm is started with a set of solutions which are represented by genes are simply known as population. Here the result from one population is taken and used to materialize a new result. The following is the simple process cycle of Genetic algorithm. The proposed system improves the classification accuracy by applying PCA results into the genetic algorithm.

- Creation of a population of sequence.
- Evolution of each sequence.
- Selection of the best sequence.
- Genetic algorithm exploits to create new results of sequence.

Algorithm: Disease diagnosis using WGA

1. Choose initial population I from PCA
2. Evaluate the fitness of each individual in the population using

$$E = \sum_{k=0}^{n} \binom{n}{k} xF^k$$

3. Repeat until termination: (time limit or sufficient fitness achieved)
   a. Select optimum -ranking individuals to reproduce
   b. Breed new generation through crossover and/or mutation (genetic operations) and give origin to offspring
   c. Evaluate the individual fitness of the offspring
   d. Replace worst ranked part of population with offspring

The details of the algorithm are given below.

1. GA generates n initial solutions using WPCA. It runs WGA for fixed number of iterations, t.

a. Choose initial population of fixed size and set j=1
b. While(j<=t)
Begin
i. Apply the operator on the two parent schedules chosen randomly to produce two offspring and replace the parents by the best two the four schedules.
ii. j=j+1
End

4. GA sends *m* best solutions chosen to the *m attributes.*
5. Each worker node runs the GA algorithm by using the initial state received.

Algorithm2: Disease diagnosis using WGA

The proposed system performs the prediction model based on the above WGA algorithm. The proposed system successfully analyses the diabetes and Heart disease based on the given training dataset. The system also predicts the score for the chance of Heart disease based on the boundary calculation. The proposed system implements a semi supervised classifier which does not depend on the training dataset completely. The system performs the statistical properties to evaluate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

IV. IMPLEMENTATION AND RESULTS

A. DATASET

Two standard clinical datasets of varying sizes and characteristics were obtained from UCI Machine Learning Repository is used in this experiment. The details of the datasets are as follows: The experiment used two datasets for diabetes. The first standard diabetes dataset from UCI Machine Learning Repository is used to discriminate healthy people from those with diabetes disease, according to class attribute which is set to either 0 for healthy and 1 for diabetes disease. This dataset contains 19 attributes and 1 categorical valued class variable and 106 records. The second data set is used to diagnose the heart disease. The dataset consist of 270 instances collected from all UCI repositories. Using some synthetic dataset a subset is used to evaluate the proposed method. We perform the experiment on the Mayo Clinic patient data obtained during the study period from 1/1999 to 12/2004 with follow-up information available until the summer of 2010. Another dataset used in this study is the Cleveland Clinic Foundation, which is named as Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 13 attributes. The experiment takes Heart disease dataset from UCI repository. The dataset contains 13 attributes considered are: age, sex, FBS (fasting blood sugar > 120 mg/dl), chol (serum cholesterol in mg/dl), restecg (resting electrocardiographic results), trestbps (resting blood pressure), thalach (maximum heart rate achieved), exang (exercise induced angina), slope (the slope of the peak exercise ST segment), oldpeak (ST depression induced by exercise relative to rest). There are a total of 750 patient records in the database. Based on the two real world dataset shown in fig 2.0, 3.0, diabetes and co morbid conditions associated with diabetes were assessed.

class	G	stab	hdl	ratio	glyhb	age	gender	height	weight	frame	bp_1s	bp_1d	Height_in_mr
Normal	0	92	37	6.1999998092...	4.6399998664...	58	1	61	256	2	190	138.04	1.549399999...
disease	0	87	0	3.5999999046...	4.8400001525...	45	0	69	166	2	160	96	1.7526
disease	0	65	0	3.0999999046...	4.6700000762...	47	0	67	230	2	137	37	1.7018
disease	0	80	57	3	6.2100000381...	57	0	71	145	1	124	64	1.8034
disease	0	85	51	3.2999999523...	6.1399998664...	40	1	65	180	1	106	82	1.651
disease	0	78	59	3	5.2300000190...	36	0	70	161	1	130	79	1.778
disease	0	69	64	2.7000000476...	4.3899998664...	20	1	64	161	1	108	70	1.6256
disease	0	84	52	3.5999999046...	5.2800002098...	53	1	61	145	1	147	72	1.549399999...
disease	0	101	36	6.8000001907...	4.6599998474...	32	0	70	212	0	0	90	1.778
Normal	7...	79	46	4	4.5900001525...	40	1	59	165	1	0	87	1.4986
Normal	8...	92	41	4.8000001907...	4.8400001525...	30	0	69	191	1	161	87	1.7526
Normal	8...	94	69	3.5999999046...	4.8099999427...	34	0	71	190	2	132	98.7	1.8034

Fig 2.0 diabetes dataset

class	age	sex	chest_pain	resting_blood_p	serum cholestoral in mg/dl	fasting blood sugar > 120 mg/dl	resting electrocardiog results	maximum heart rate achieved	exercise induced angina
Heart_di...	70	1	4	130	322	0	2	109	0
Normal	67	0	3	115	564	0	2	160	0
Heart_di...	57	1	2	124	261	0	0	141	0
Normal	64	1	4	128	263	0	0	105	1
Normal	74	0	2	120	269	0	2	121	1
Normal	65	1	4	120	177	0	0	140	0
Heart_di...	56	1	3	130	256	1	2	142	1
Heart_di...	59	1	4	110	239	0	2	142	1
Heart_di...	60	1	4	140	293	0	2	170	0
Heart_di...	63	0	4	150	407	0	2	154	0

Fig 3.0 Heart dataset

In this paper, Principal Component Analysis is used for Feature extraction and Weighted Genetic Algorithm (WGA), a supervised learning method is used for disease diagnosis. Feature extraction transforms the data in the high-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

dimensional space to a space of fewer dimensions. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. The huge size in data may affect the algorithm performance. So, Feature extraction is a mandatory process for constructing combinations of the variables to get around these problems. Best results are achieved when the features are constructed from the effective feature selection process and a set of application dependent features. Feature extraction is implemented using the Principal Component Analysis method and Linear Discriminate Analysis.

Weighted Genetic Algorithm has been successfully applied for various pattern recognition problems. It is primarily a classifier that performs classification tasks by constructing an optimal chromosome selection in a multidimensional space that separates members and non-members of a given class.WGA is attractive for clinical data analysis for its efficiency in handling sample sizes in the order of millions, its flexibility in choosing a similarity function and its ability to identify outliers.

DATA SIZE	GA	WGA
100	0.4	0.098
200	0.3	0.105
300	0.45	0.13
450	0.48	0.12
500	0.5	0.3

Table 1.0 False Positive Rate comparison table

From the results shown in Table 1.0, it is found that WGA Based Classification Optimization increase the True Positive Rate significantly and reduce the False Positive Rate to an acceptable extent.

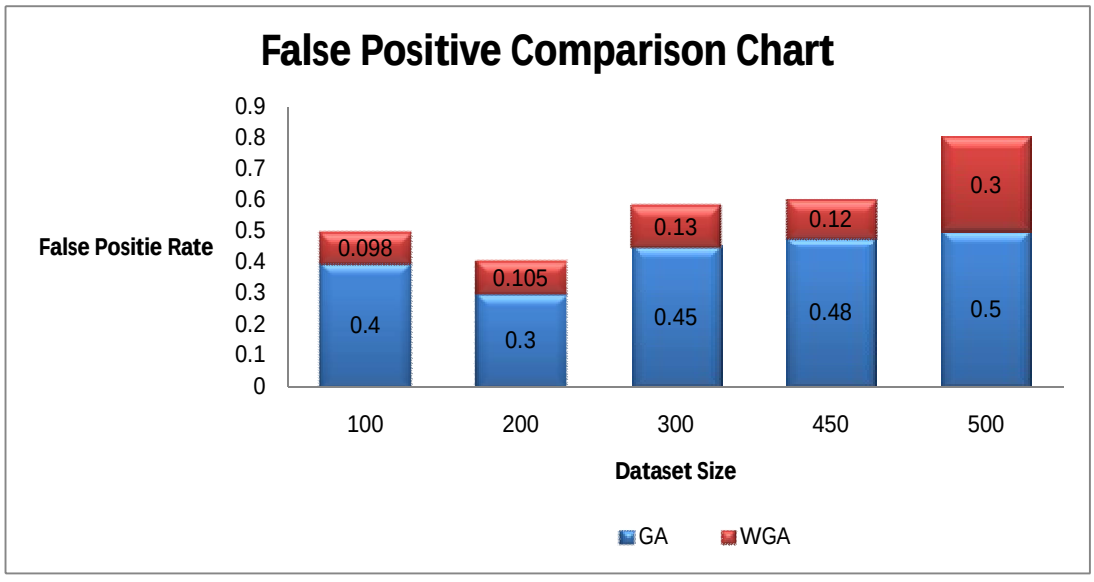

Figure 4.0 Performance Measure in terms of False Positive Rate

As in fig 4.0 The false positive rate of the proposed system is quite high, because some normal classes in the additional data merged could be clustered as a disease, but only the weighted features are used in grouping. The reduction in false positive rate of the proposed system is mainly due to the WGA and genetic process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

VI. CONCLUSION

In this paper, we proposed a new classification and prediction scheme for diabetes and Heart disease data. The system studied the main two problems in the literature, which are diagnosis accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced weighted component with genetic algorithm. The PCA represents with the effective splitting criteria which has been verified by the genetic algorithm. The system effectively identifies the disease and its sub types, the sub type which is referred as the percentage of class such as normal and disease.

The experimental results are evaluated using the C#.net. The experimental result shows that integrated extended weighted component with genetic algorithm shows better quality assessment compared to traditional GA techniques. From the experimental results, the execution time calculated for classification object is almost reduced than the existing system. The proposed framework model can be used to analyse the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for classification performance.

REFERENCES

[1]. I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[2]. G. D. Magoulas and A. Prentza, "Machine learning in medical applications," *Mach. Learning Appl. (Lecture Notes Comput. Sci.)*, Berlin/Heidelberg, Germany: Springer, vol. 2049, pp. 300–307, 2001.
[3]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.
[4]. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.
[5]. Wang, Xuechuan, and Kuldip K. Paliwal. "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition." *Pattern recognition* 36.10 (2003): 2429-2439.
[6]. Ruben, D.C.J., *Data Mining in Healthcare: Current Applications and Issues.* 2009.
[7]. Porter, T. and B. Green, *Identifying Diabetic Patients: A Data Mining Approach.* Americas Conference on Information Systems, 2009.
[8]. Panzarasa, S., et al., *Data mining techniques for analyzing stroke care processes.* Proceedings of the 13th World Congress on Medical Informatics, 2010.
[9]. Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, *Data mining techniques for cancer detection using serum proteomic profiling.* Artificial Intelligence in Medicine, Elsevier, 2004.
[10]. Das, R., I. Turkoglu, and A. Sengur, *Effective diagnosis of heart disease through neural networks ensembles.* Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.