International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4s, November 2018

Heart Disease with Risk Prediction using Machine Learning Algorithms

S. Kavitha, K. R. Baskaran, S. Sathyavathi

Abstract: ***Nowadays health connected issues are terribly high, and it can'tbe simply foretold earlier to avoid complications. Wellness heart condition cardiopathycar diovascular disease} (HD) may be a common disease for the individuals more matured cluster thirty-five to fifty. the sector of information mining has concerned within the medical domain, With the historical knowledge, mining algorithms are able to predict and classify the abnormality in conjunction with its risk levels. The previous studies related to predict heart problems have used several features which has been collected from patients. The accuracy level of prediction and the number of features is very less in the previous systems. To improve the prediction accuracy the planned system, consider additional range of options and implements a Weighted Principle Analysis (WPCA) and changed Genetic Algorithm(GA) The planned technique helps the medical domain for predicting HD with its numerous co-morbid (types of heart diseases) conditions. The system has 2 main objectives, that are rising diagnosing accuracy and reducing classification delay. The WPCA represents with the effective cacophonous criteria that has been applied into the genetic Algorithm. The system effectively identifies the disease and its sub types, the sub type which is referred as the level of class such as normal and mild or extreme.Using combinatorial methods from data mining decision making has been simplified and the proposed work achieved 96.34% accuracy, which is higher than the known approaches in the literature.***

Keywords: Data mining, Classification, Weighted Principle Analysis (WPCA), Modified Genetic algorithm (GA), Heart Disease.

I. INTRODUCTION

Blood pressure, cholesterol, pulse rate, stress, food habits are the factors that contribute to the Heart disease.The main functioning organ of the human body is heart.If the human heart isn't functioning it will affect the entire human body. Some risk factors of cardiopathy organ are family background, stress level, cholesterol level, Age, food diet. Inhaling of tobacco. Blood vessels unit overstretched indicates the danger level of the body pressure.The pressure at middle muscle of the heart, exaggerate the level of lipids overthe time in the blood causes heart disease. Lipids settle in the arteries and block the flow of blood to the cardiac organ..

Manuscript published on 30 November 2018.
*****Correspondence Author(s)
S. Kavitha, Assistant Professor, Assistant Professor, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.
K. R. Baskaran, Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.
S.Sathyavathi, Assistant Professor, Department of Information Technology. Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.

Inhaling the tobacco is the root cause for cardiac arrest that ends up in death. As the pressure increases in thicken the blood. The Naïve mathematician techniqueis employed to predict the center illness through likelihood. The Neural Network provides the reduced error of the prediction of cardiopathy. By monitoring the activity of the patients continuously the death rate can be reduced..

The model will help the doctors to diagnose the disease earlier engineering This paper projected a brand new combinatorial technique to search out the danger of cardiopathy. This additionally aims to develop a high accuracy classifier for top dimensional datasets.

II. EXISTING SYSTEM

There are many approaches within the literature for detecting the arteria coronaries illness [1]. within the existing work, decision tree classification algorithm has been to assess the events associated with cardiac disease[3]. In decision tree ID3, C4.5 algorithms are used to perform CAD analysis. Under classification stream, Support Vector Machines (SVM) and fuzzy logics are used.

Existing studies do not include more features of heart disease. In this paper, the more features are considered to improve the diagnosis accuracy. Adding new discriminative features to the patients' records has an effect on prediction efficiency of the algorithms.

Table 1.0 Existing Algorithms:

Algorithm	Type of data	Speed	Pruning	Missing Values
PCA algorithm	Continuous	low	yes	Can't deal with
ID3	Categorical	Low	No	Can't deal with
C4.5	Continuous and Categorical	Faster than ID3	Pre-pruning	Can't deal with

III. PROPOSED SYSTEM

The chapter discusses about the proposed methodology and the steps involved in that. The proposed system uses WPCA for feature extraction and GA to improve the prediction accuracy.The following are the main contributions of the proposed work.

- The system implements a new Genetic based algorithm with the use of effective risk prediction. The system introduces a new Heart Disease Classification algorithm with GA technique.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

- This also creates a new fusion approach for fast disease classification by using feature
- selection and genetic algorithms. The system developed with the intension of high accuracy and less training overhead.
- So, the system initially collects and make score for every label, this partially makes an ensemble approach to improve the detection speed.
- WPCA for feature selection and dimensionality reduction
- Genetic algorithm for disease classification and prediction.

The pruned data has been applied in the formula of WPCA formula has results in the new best classification algorithms. It might handle giant class dataset with better accurately and efficiency.

IV. METHODOLOGY

A. Modules:

Data set collection and uploading

- Heart disease dataset from UCI repository is ued.

I. Preprocessing
- Handle missing valueson the pre-processed data and select best features. values by calculating mean and other statistical process

II. Feature selection using WPCA
- WPCA algorithm implementation

III. Genetic algorithm implementation
- Apply the result of WPCA into the genetic algorithm selection section
- Selection
- Cross over
- mutation

IV. Heart disease classification with co-morbid conditions and severity level
- Training process
- Test process
- Heart disease predicted report with the sub type of disease and the score

V. Alert module
- Alerts using SMS or email/sound alert if the disease predicted

B. DATASET

Cleveland dataset, which is collected from UCI repository. The system uses 303 datasets for the evaluation. The system contains the following sub tasks.

a. Training the dataset

Preprocessing (handling the missing values, noisydata) is the method of eliminating the inconsistent data.Classification algorithm will handle the redundant record sets. Here are not any duplicate records among the planned check sets; therefore, the performance of the learners is not biased by the ways that have higher detection rates on the frequent records.

b. Initial clustering

The Chunk is that the set of information sorted along. The system implements the Chunk based mostly techniques for cluster. This module describes the "Chunk system" methodology supported such character of Chunks, that received nice experimental laboratory results. This module implements the Chunk cluster method.
After third clustering, marked objects neighbors will be stored in training data file which can be accessed by component feature selection algorithm; which has been resulted with better robustness, positive feedback along with distributed computing methods.

C. Extraction of Features.

WPCA steps

- Taking the whole dataset. ignoring the class labels
- Find initial component
- Calculate the d-dimensional mean vector
- Calculate the covariance matrix of the original or standardized d-dimensional dataset X (here: d=3); alternatively, compute the correlation matrix.
- Eigen decomposition: Calculate the eigenvectors and Eigen values of the covariance matrix (or correlation matrix).
- Arrange the Eigen values in descending order.
- Choose the k eigenvectors that correspond to the k largest Eigen values where k is the number of dimensions of the new feature subspace ($k \leq d$).
- Calculatethe matrix of projection W from the identified sample k.
- Convert theactual dataset X to obtain the k samples dimensional feature subspace Y

It establishes the improved model of three hundred (Constrained Co Clustering)-based approach, that may be a hybrid of the constraint approach and therefore the feature choice approach. By continuation the processes of CONSTRAINT coaching and co cluster, the detection feature is established and keep within the common storage Disk, that is employed within the testing part. this may finally have accustomed show the results.

D. Hellinger distance and oversampling implementation

In this module, a new notion of distance between probability distributions called Hellinger distance is used.

a. WPCA+GA setup:

The investigation knowledge has threehundred perceptive sample, there exists missing worth in these sample. once eliminating the missing part, the system performs the weighted principle component analysis and Genetic algorithm for each attribute. The weighted principle component analysis and Genetic algorithm implementation method identifies the frequency of each worth from the dataset. Unlike previous algorithms, this doesn't store the complete knowledge matrix or variance matrix, and therefore the approach is particularly of interest in dynamic or large-scale dataset. gearedtoward the vital influence of weighted principle component analysis and Genetic algorithm with Chunk primary direction on classification performance, this paper adopts theweighted principle component analysis and Genetic algorithm because the technique of choice optimum characteristics parameters.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

b. **Classification**:

weighted principle component analysis and Genetic algorithm based mostly classification has been created during this module. The user will offer the partition threshold. a collectionknowledge|of knowledge of information} instances within the original data set is taken as predefined input. This knowledge is also contaminated by noise and incorrect knowledge labelling etc., this knowledge may be error free, because of this can beprogressing to be used as coaching knowledge. So, the cleanup is completed victimization before change the information. This has been applied by oversampling methodology.

c. Test results

once the user offers the input to the system, the system performs the weighted principle component analysis and Genetic algorithm and Oversampling worth for the new input. Compare new sample S_t with the trained data and it help to predict.

d. Performance results

The module results in predicting the disease earlier with high accuracy and the efficiency.

V. IMPLEMENTATION OF THEMODEL

Dataset: Two standard medical datasets that vary in their characteristics has been obtained from UCI Machine Learning Repository which has been implemented in this experiment. The experiment used 2 datasets for polygenic disease. the primary commonplace wellness} dataset from UCI Machine Learning Repository is employed to discriminate healthy individuals from those with polygenic disease disease, in line with category attribute that is about to either zero for healthy and one for wellness} disease. This dataset contains nineteen attributes and one categorical valued category variable and 106 records. The second information set is employed to diagnose the center wellnessThe dataset consists 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 dressing Clinic patient knowledge obtained throughout the study amount from 1/1999 to 12/2004 with follow-up info accessible till the summer of 2010. Another dataset employed in this study is that the Cleveland Clinic Foundation, that is known as as heart condition knowledge set accessible at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. the info set has thirteen attributes. The experiment takes heart condition dataset from UCI repository. The dataset contains thirteen attributes thought of are: age, sex, FBS (fasting blood glucose > one hundred twenty mg/dl), chol (serum steroid alcohol in mg/dl), restecg (resting medical instrument results), trestbps (resting blood pressure), thalach (maximum rate achieved), exang (exercise evoked angina), slope (the slope of the height exercise ST segment), recent peak (ST depression evoked by exercise relative to rest). There area unit a complete of 750 patient records within the information. supported the two-real world dataset, were assessed.

class	age	sex	chest_pain	resting_blood_	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

Fig1 Heart dataset

VI. RESULT

The experiments square measure designed so the various components of the work can be evaluated. The WPCA+GA square measure enforced victimization C#.net. All the four potential combos of the feature choice and creation strategies square measure in theory analyzed over the dataset. Theme was compared with the prevailing algorithms supported the subsequent parameters.

- Specificity –estimate the proportion of negatives that is identified correctly.
- Sensitivity- estimates the proportion of positives that is identified correctly
- Accuracy – Determines the correctness
- Precision –Repeated process same result
- Time taken – Determines the processing time involved.

Sensitivity, specificity and accuracy are described asTrue Positive, True Negative, False Positive and False Negative.

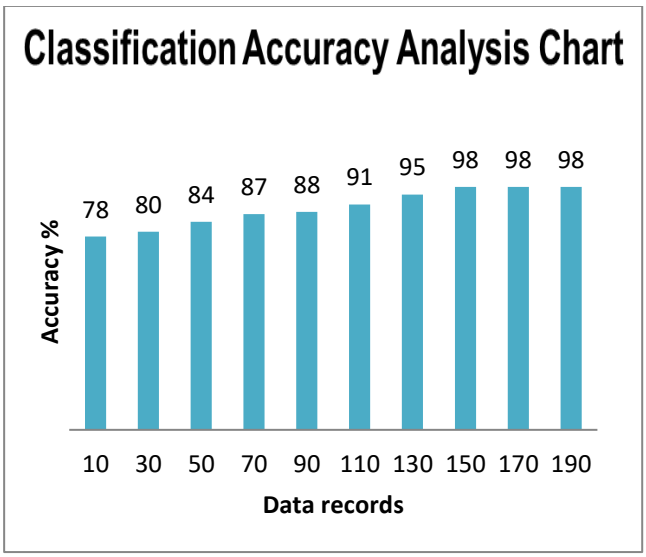

Fig 2 Classification Accuracy analysis chart

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

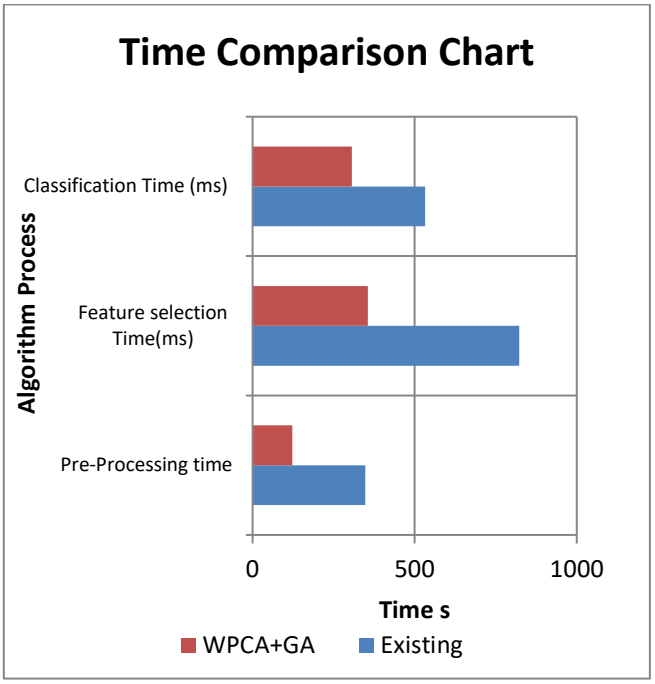

Fig 3 Time Comparison Chart

VII. CONCLUSION AND FUTURE WORK

The study planned a brand-new classification and over sampling technique to scale back the category imbalance downside and classification accuracy issues. The system studied the most 2 issues within the literature, that square measure spatiality and classification accuracy. The study overcomes the on top of 2 issues by applying the effective increased WPCA+GA with Hellinger distance calculation. The experimental results square measure evaluated mistreatment the C#.net. The experimental result shows that planned system could yield higher quality assessment compared to ancient oversampling and classification techniques. From the experimental results, the execution time calculated for classification object is sort of reduced than the present system.

VIII. FUTURE WORK

The projected framework model is often wont to analyze the prevailing work, establish gaps and supply scope for more works. The analyzers could use the model to spot the prevailing space of research within the field of information mining in alternative dataset and use of alternative classification algorithms. As more work, use this model as a useful base to develop associate degree acceptable data processing system for classification performance.

REFERENCES

1. Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*. IEEE, 2016.
2. Wilson, Peter WF, et al. "Prediction of coronary heart disease using risk factor categories." *Circulation* 97.18 (1998): 1837-1847.
3. http://www.americanbusinessmag.com/2010/04/symptoms-of-a-heart-attack-2/
4. Palaniappan, Sellappan, and RafiahAwang. "Intelligent heart disease prediction system using data mining techniques." *Computer Systems and Applications, 2008. AICCSA 2008.IEEE/ACS International Conference on*.IEEE, 2008.
5. Khaing, HninWint. "Data mining based fragmentation and prediction of medical data." *Computer Research and Development (ICCRD), 2011 3rd International Conference on*. Vol. 2. IEEE, 2011.
6. Patel, Ajad, Sonali Gandhi, SwethaShetty, and BhanuTekwani. "Heart Disease Prediction Using Data Mining." (2017).
7. Wghmode, MrAmol A., MrDarpanSawant, and Deven D. Ketkar. "Heart Disease Prediction Using Data Mining Techniques." *Heart Disease* (2017).
8. http://archive.ics.uci.edu/ml/datasets/Heart+Disease
9. S.Kanagaraj, N.Rajathi, R.Brahmanambika, K.Manjubarkavi,, "Early Detection of Dengue Using Machine Learning Algorithms" " International Journal of Pure and Applied Mathematics " , Special Issue, February, 2018.
10. K.R. Baskaran, V. Vijilesh, R. Nedunchezhian and R.S. Kumar "Combining Intelligent Web Caching with Web Pre-Fetching Techniques To Predict Tourist Places ", International Journal of Pure and Applied Mathematics " Volume 116 No. 12 2017, 97-105
11. K.R.Baskaran, C. Kalaiarasan, "Improved Performance By Combining Web Pre-Fetching Using Clustering With Web Caching Based On Svm Machine Learning Method", International Journal of Computers Communications & Control, ISSN 1841-9836,Vol.11, No.2, April 2016, pp. 166-177
12. K. R. Baskaran, C. Kalaiarasan, "Pre-Eminence Of Combined Web Pre-Fetching And Web Caching Based On Machine Learning Technique", Arabian Journal for Science and Engineering (Springer Journal), ISSN 1319-8025, Vol. 39, No.11, November 2014, pp. 7895-7906

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)