A NOVEL MORTALITY PREDICTION APPROACH IN CONGESTIVE HEART FAILURE PATIENTS USING RF-IWFFO

C.Sowmiya[1], Dr.P. Sumitra[2]
[1]Ph.D Research Scholar, [2]Assistant Professor,
PG and Research Department of Computer Science and Applications
[1,2]Vivekananda College of Arts and Sciences for Women (Autonomous),
Elayampalayam, Tiruchengode-637205,TamilNadu,India

Abstract:
In modern society, Heart disease is the noteworthy reason for short life. Large population of people depends on the healthcare system so that they can get accurate result in less time. Large amount of data is produced and collected by the healthcare organization on the daily basis. To get intriguing knowledge, data innovation permits to extract the data through automization of processes. In this research, a Random Forest algorithm with Intensity Weighted FireflyOptimization (RF-IWFFO) is proposed for heart disease prediction. The performance evaluation of the proposed system is compared with the prior SVM with Recursive Feature Elimination.
Index Terms: Heart Disease, Random Forest, Congestive Heart Failure FireFly Optimization

I. INTRODUCTION

The Process of finding knowledge and information from a vast database is known as data mining. To turn large amount of data into useful information, data mining is used [1]. Association rules give the relationship between the items that are present in large database and are a vital way of knowledge representation. As Association rule is becoming one of the most researched area, the database community is now giving more attention to the association rule. Association rule mining was given by A.Swan, T.Imiclinski and R.Agrawal [2]. Due to large amount of data and using that data to extract useful information is the major reason that data mining has attracted huge attention in recent years in information industry.

Due to heart disease almost 23.6 million people will die in 2030 as estimated by World Health Organization. The analysis of coronary illness relies upon clinical information. When the clinical data of a patient is present, then the heart disease prediction system can help in predicting coronary disease accurately. The healthcare industry is collecting data of patients in large amount which can be mined to discover hidden information that can help medical professionals in effective decision making. There are many reasons for heart disease such as stress, high blood pressure, drug abuse, lack of exercise, food habit, cholesterol, etc. Our blood vessel becomes weak due to fatty food which can lead to heart disease.

The walls in heart become thicker when more pressure is applied to our arteries. As the walls become thick, it can slow down the flow of blood and can also make the block which lead to heart disease [3] [4]. So we are introducing a method for predicting the heart disease. The pattern that appear frequently in a dataset are called frequent pattern. To find interesting patterns from large database, frequent item set play an important part in information mining. The records of crores of people can be stored and also the information about their treatment. These along with the data mining strategies can help in answer the most important questions which are related to health of a patient [5]. This paper is roused by the perspective and the previously mentioned issues and proposes an arrangement of methodologies for heart disease prediction.

In this paper a firefly algorithm is utilized for feature selection and the random forest is applied for classification approach. The performance evaluation of this result is compared with the prior approach performed with recursive feature elimination and SVM.

II. Related works

This section describes the literature performed by researchers regarding Congestive Heart Failure prediction using Machine Learning (ML) techniques.

Shouman et al. [6] diagnose the cardiovascular disease with high accuracy but it is not easy to achieve it. Additionally, a combination of significant features will definitely improve the accuracy of prediction. This shows that an extensive experiment to identify significant features is necessary to achieve that goal.

Cheng-HsiungWenget al [7]analyzed the different classifiers, including an ensemble classifier and solo classifiers. Further, researcher uses various evaluation factors to evaluate the performance of these classifiers with real-life datasets. Eventually, a statistical testing is used to evaluate the importance of the changes in performance among the three classifiers.

Nahar et al. [8]presentsa proper evaluation and comparison to test the different combination of features together with the data mining techniques is yet to be focused. Thus, a proper experimentation is required to provide proper identification of data mining approaches and relevant features to ensure the prediction of heart disease is accurate.

Dey et al., [9] found the best combination of important features that works well with the best performing algorithm. This studyaims on discovering the data mining techniques with required features that will perform well in heart disease prediction. Even though, it is not easy to discover the proper technique and choose the necessary features.

Kavitha and Kannan [10]discover that data redundancy in a raw dataset affect the predicted result. Likewise, in order to use the machine learning algorithms to its full potential, aproper preparation is required for preprocess the datasets.

III Prior Prediction Approach

The prior approach utilizes the recursive feature elimination and SVM methods for heart disease prediction.

SVM-RFE

Recursive Feature Elimination (RFE) is apowerful procedurein feature selection that depends on the specific learning model. Initially RFEis proposed by Guyon et al. [20] for cancer classification by using SVM. RFE use all features to build a SVM model at first, and then ranks each feature contribution in the SVM model thatproduces a ranked feature list, and finally removesunwanted features in the SVM model. To compute the contribution of each feature, RFE uses $\|w\|^2$was the ranking criterion, where w is a weight vector. Hence, the unwanted features are found based on whose value of $\|w\|^2$ is small. Such that, the features with highest values of $\|w\|^2$w are selected and the features with lower values are removed.

In SVM [21], w in the feature space can be written as

$\|w\|^2 = \left|\sum_{i=1}^{n} \alpha_i y_i \emptyset(x_{ij})\right|^2$, j=1,2,…..m (1)

where$\emptyset(x_{ij})$means the i -th instance of j -th feature vector.

The procedure of RFE is as follows:

RFE Algorithm

1) Build the SVM model using all candidate features-

2) Evaluate the contribution of current features to SVM by calculating ranking criterion with current features using (2);

3) Rank current features according to their contributions (a high value of ranking score stands for great contribution) and produce a ranked feature list;

4) Remove a specific number of features at the bottom of the ordered list, and use the rest features as the new candidate features;

5) Go to the step 1), until it satisfied the specific number of features.

III. Proposed Methodology

The proposed methodology for Heart Disease using data mining process is described in this section. The architecture of the study is shown in figure 1. Initially the very first step is preprocessing the data. Preprocessing is the process of missing value estimationbased on categorical or nominal data. Further the selection of the best features is performed byIntensity WeightedFirefly Optimization. Then the classification process is performed to predict Congestive Heart Failure.

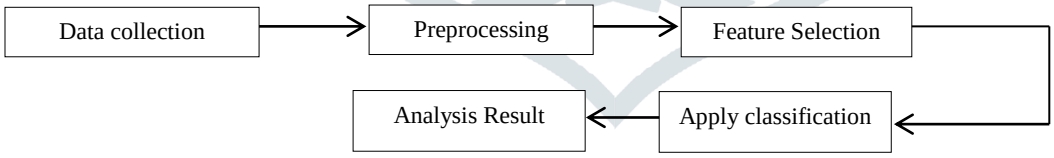

Figure 1. Architecture

Intensity Weighted Firefly Optimization

Feature selection is process of selecting useful feature from dataset. Fire fly is a kind of a flash light which tries to communicate with the other members of their nature. As the intensity of light vanishes with respect to distant locations, its accuracy can be defined at local horizons for finding the best solution for any function. In this article, the fire flies are the particles or the extracted features from peak estimations. Each extracted feature (fire fly) is assigned by light intensity and Out of all the extracted features, the distinct features which have common species are selected as the best one. This is best explained by the contours in which random regions are created based on the nature of features extracted and the particles of similar species are attracted towards the centre of the regions of the contours.

The random regions are created based on the feature categories and the particles of similar nature will follow their own regions. Out of all the particles, some are in the centre of the regions and these are defined as the best features for better classification of disease. Hence, fire fly optimization will serve the purpose of feature reduction technique by considering similar natured particles and neglecting the others.

Let us consider T_F as the feature vector or feature matrix. On selecting a training feature, T_r ' Define $\propto, \beta \; and \; \gamma$ with some random values (here 0.2, l.0 and l.0 are considered respectively).

Let $X = X_i \; (i = 1,2,3 \ldots\ldots\ldots n)$-
(2)

Where 'n' is the number of particles and 'X' is the population of fire flies.

$Define \; I = rand(T_F)$-
(3)

Where I is the light intensity. Updating the observation coefficient as

$$\gamma_i = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$
(4)

Where i=I,2, 3, ... n, j=I,2, 3, ... m. Final updates are expressed as

$$x_n = x_n(i) * (1 - \beta) + x_n(j) * \beta + \alpha(\text{rand} - 0.5) \quad (5)$$

$$y_n = y_n(i) * (1 - \beta) + y_n(j) * \beta + \alpha(\text{rand} - 0.5)$$
(6)

When the light intensity gets updated after some iteration, the final values are indicated as

$f_t = I\,(x, y)$ ⟶ Exact fitness value

$idx = \min(f_t)$ ⟶ Exact best fitness value

$T_r = T_F\,(idx)$ ⟶ Selected best feature

Random forest Classifier:

Random forests are an ensemble method that combines multiple independently trained decision trees. Each tree is constructed using bootstrap samples from the training data. For each split in the tree, feature selection is performed from a small, randomly selected subset of the feature space. In contrast to training in general decision trees, which requires post-pruning, the trees in random forests do not need pruning because random sampling guarantees no over-fitting. The final prediction of random forests is an average of or majority vote from the predictions of all trees.

Over fitting which exists in traditional decision tree can be avoid in Random Forest. Random Forest classifier is the algorithm consisting of a combination of tree classifiers introduced by Breiman and Cutler[10]. The 10-fold cross validation results were selected as the standard by which to compare Random Forest results with other models. By bootstrap resampling, several tree classifiers are built. Through the K round of training, a classification model series acquired, $\{k\,\{h_1\,(x), h_2\,(x),, h_3\,(x)\}\,k\,, h_4\,(x)\}$, then they are used to construct classifiers, the final result of the Random Forest system is passed by simple majority vote. The final decision is:

$$H\,x = argmax \sum_{i=1}^{k} I(h_i \; x = Y)$$
(7)

H(x) is the classification model of combination, and h(x) is single decision tree classifier. Y is output variable. I (.) is Indication function[12].

IvExperimental Result

The performance evaluation of the study is discussed in this section.The proposed system is implemented in java language. The NetBeans IDE is utilized for front end design. MYSQL 5.1 is used for database access. The parameters such as Precision, Recall, F-Measure and its Accuracy are used in this approach to evaluate the performance of the prediction process. The metrics are measured by applying the classification technique. The parameter precision and recall is calculated to obtain the correctly classified class. Precision is computed to find the presence of relevant class and it is the fraction of total classified class that is relevant.

The formula for precision calculation is

$$\text{Precision} = \frac{\text{n}(\{\text{correctly classified class}\} \cap \{\text{total classified class}\})}{\text{n}(\text{total classified class})} * 100\% \quad (8)$$

Recall is calculated to find selected relevant class and it is the fraction of correctly classified review sentence from the total classified sentence. The formula for precision calculation is

$$\text{Recall} = \frac{n(\{\text{correctly classified class}\} \cap \{\text{total classified class}\})}{n(\text{correctly classified class})} * 100\% \quad (9)$$

F-Measure is mainly used in information retrieval field to find the test accuracy. F-measure is the weighted average of precision and recall. The f-measure can be computed by following formula

$$\text{Fmeasure} = 2 * \frac{\text{precision} * \text{recal}}{\text{precision} + \text{recall}}$$

Table 2. Comparison Result of SVM-RFE with Proposed System

Algorithm	Precision	Recall	F-Measure	Accuracy
SVM-RFE	90.3%	89.4%	94.5%	94.5%
RF-IWFFO	93.2%	90.1%	98.7%	98.7%

Table 2 describes the performance of theproposed methodology. The SVM-RFE obtains 94.5 percent accuracy whereas RF_IWFFO obtain 98.5 % accuracy which shows that the Proposed approach obtain a better accuracy than the previous method.

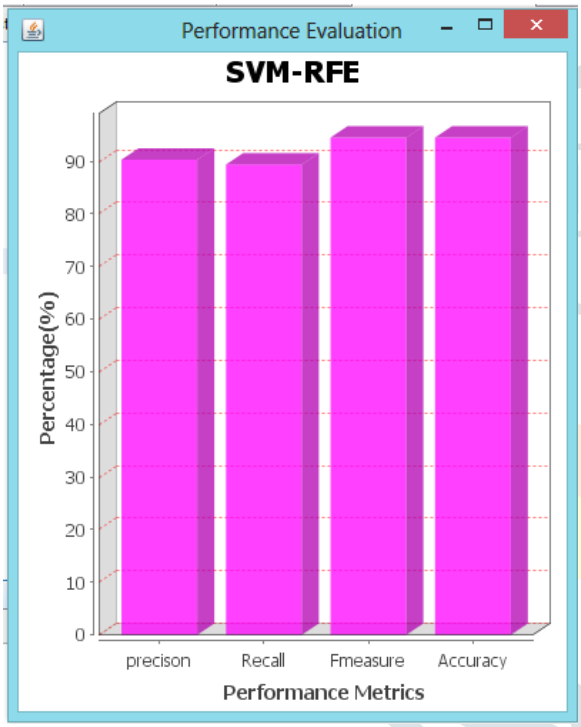

Figure 2. Result of SVM-RFE

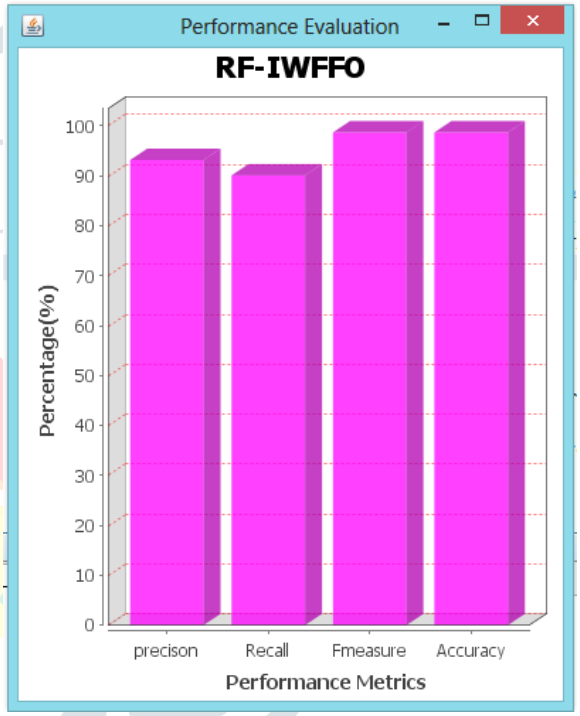

Figure 3. Result of RF-IWFFO

V Conclusion

Data mining is the process of extracting useful and previously unknown patterns from huge database or data warehouse. Now a day's data mining plays important role in many sectors some of this are in health sector, Bank, financial sector, education sector etc. Different researches have been done about a prediction of heart using different algorithm. This study is intended to predict Congestive Heart Failure using classification algorithm. SVM-RFE algorithm gives 94.5% predictive accuracy, whereas Random forest-IWFFO algorithm gives 98.7% predictive accuracy. These results were obtained with less number of iteration and it shows improvement from SVM-RFE paper

.

References:

[1]. Aswathy Wilson, Gloria Wilson, Likhiya Joy, "Heart Disease Prediction Using the Data mining Techniques", International Journal of Computer Science Trends and Technology, Jan-Feb 2014, pp. 84- 88.

[2]. ShashiChhikara, Pururshottam Sharma, "Weighted Association Rule Mining: A Survey", International Journal for Research in Applied Science and Engineering technology, 2014, pp. 85-88.

[3]. Aswathy Wilson, Gloria Wilson, Likhiya Joy, "Heart Disease Prediction Using the Data mining Techniques", International Journal of Computer Science Trends and Technology, Jan-Feb 2014, pp. 84- 88.

[4]. Purushottam Sharma, DrKanakSaxena, Richa Sharma" Efficient Heart Disease Prediction System using Decision Tree" in IEEE International Conference on Computing Communication and Automation (ICCCA-2015),May 2015.

[5]. Feng Tao, fionnMurtagh, Mohsen Farid, "Weighted Association Rule Mining using Weighted Support and significance Framework", 2003.

[6]. Shouman, M., Turner, T., Stocker, R. 2011. Using decision tree for diagnosing heart disease patients. Proceedings of the Ninth Australasian Data Mining Conference (AusDM'11), Darlinghurst, Australia, Volume 121, pp. 23-30

[7]. Weng, C.H., Huang, T.C.K. and Han, R.P., 2016. Disease prediction with different types of neural network classifiers. *Telematics and Informatics*, *33*(2), pp.277-292.

[8]. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P., 2013. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications. 40(1), 96-104.

[9]. Dey, A., Singh, J., Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. Analysis. 140(2), 27-31.

[10]. Kavitha, R., &Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, pp. 1-5.

[11]. Breiman L. Random Forest[J]. Machine Learning, 2001, 45:5- 32.Breiman L. Random Forest[J]. Machine Learning, 2001, 45:5-32

[12]. Fang K N, Jian-Bina W U, Zhu J P, et al. A Review of Technologies on Random Forests[J]. Statistics & Information Forum, 2011.