ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse72952020.pdf
https://doi.org/10.30534/ijatcse/2020/72952020

An Ensemble Feature Selection Method for Prediction of Chronic Diseases

Manonmani M[1], Dr. Sarojini Balakrishnan[2]
[1]Avinashilingam Institute for Home Science and Higher Education for Women, India, manonmaniatcbe@gmail.com
[2]Avinashilingam Institute for Home Science and Higher Education for Women, India, dr.b.sarojini@gmail.com

ABSTRACT
Feature Selection is an important preprocessing step in data mining to provide enhanced results in the medical domain. The proposed research work aims to select the most discriminatory features in chronic medical datasets. The proposed method (D-ITLBO) is an ensemble method that combines a filter and wrpper approach to five medical datasets. Firstly, DFS method applies Probability Density Function (PDF) to each feature and ranks them. Then, the features selected from DFS method are applied to wrapper-based optimized feature selection algorithm known as Improved Teaching Learning Based Optimization (ITLBO) algorithm for finding the optimal feature subset. The derived optimal feature subset consists of the most relevant and significant features that provide important information about patients for prediction of chronic illness. The results of the proposed method are evaluated using Support Vector Machine – Radial Basis Function (SVM-RBF), Gradient Boosting and Convolution Neural Network (CNN) classifiers. Empirical results indicate that the proposed methodology has achieved significant feature reduction and classification accuracy for the five medical datasets taken for evaluation.

Key words : Convolution Neural Network, Ensemble Feature Selection, Gradient Boosting, Probability Density Function.

1. INTRODUCTION

Chronic Illness is a category of disease that lasts for a long duration in a patient's life. There are many chronic diseases like diabetes mellitus, kidney disease, hepatitis, thyroid, Parkinsons, coronary artery disease etc. which have to be detected at an early stage and have to be controlled in the long run to maintain the patients' health status. The advancements in AI and Machine Learning make it possible to have deep insights into the data in the medical databases. Clinical decision support systems play an important role in assisting the physicians to make a better clinical diagnosis [1]. Medical diagnosis is viewed as a significant yet convoluted task that should be executed precisely and proficiently [2]. Accurate diagnosis and the early prediction of chronic diseases can save many people who are suffering from long term illness. The large corpus of heterogeneous medical datasets may contain irrelevant and redundant features. Not all the features in the data may be required for diagnosis. These healthcare data are to be analyzed to extract patterns for effective diagnosis. The analytical relations between the input and output data also has a significant effect in determining the attributes that are required for diagnosis. The identification of features that are required for accurate diagnosis is carried out by the process of Feature selection, one of the preprocessing techniques. Feature selection approaches helps in reducing the dimensionality of the dataset by the removal of irrelevant, redundant attributes from the dataset which leads to the formation of a feature subset with most significant features that are vital in prediction of the diseases. Research show that the features subset with relevant features could also enhance the accuracy of prediction [3],[4]. In recent literatures, another important issue that is given attention is about the stability of the feature selection subsets i.e., the derived feature subset should remain insensitive to the variation in the training set [5]. When feature selection methods are adopted to identify the critical attributes that identify the disease, the stability of the results is very important. The feature selection results depend on the variations in the training data and testing data for different datasets. When the clinical data are updated in the existing database, the FS methods have to consider this dynamic nature of data entry and produce results that prove to be robust for any change in the existing performance of the classification algorithms. If such consideration is taken care of, the issue of instability that otherwise dampens the confidence of the data analysts in experimental analysis of the feature subset can indeed be overcome. Ensemble feature selection can be used to obtain a feature subset that increases the stability of the feature selection algorithm [6].

The ensemble method integrates two or more feature selection techniques to obtain the final feature subset. Ensemble techniques are highly beneficial in medical domain as through more than one approach the feature importance with respect to the target task is determined. The results of the first

stage of the feature selection method is given as the input to the second stage and so on and finally aggregates the results as a final (ensemble) result [5], [8].

In this research work, the proposed ensemble feature selection technique combines a filter-based feature selection technique, and an optimization algorithm to derive stable feature subsets. The features are ranked based on the estimates of their density value and the feature subset with features that are above a threshold value are given as the input to the wrapper based ITLBO algorithm. The measures that indicate the performance of the classifiers used find out the effects of the results of the optimal feature subset. Five benchmark medical datasets are used in this research work [8]. These datasets are found in the UCI repository.

2. RELATED WORKS

Different algorithms and methods have been proposed and implemented for prediction of chronic illness by mining the medical data of the patients.

V. R. Elgin Christo H. Khanna Nehemiah, B. Minu and A. Kannan (2019) have proposed a framework for diagnosis of Breast Cancer and Hepatitis disease by employing bio-inspired algorithmsand back propagation neural network classification for feature selection and classification respectively. Missing values in the datasets were handled using Hot-deck computation method and data was normalized using Min-Max normalization. Fitness function was calculated by finding the accuracy of AdaBoostSVM classifier. From the results derived from the application of the bio-inspired algorithm on feature selection process, Optimal feature set is derived from correlation-based ensemble method. The authors were able to achieve 98.47% and 95.51 % accuracy for WDBC dataset and Hepatitis dataset respectively.

E. Emary, H. M. Zawbaa, and A. E. Hassanien (2016) have used Binary Grey Wolf Optimization for feature selection. Empirical results were derived from the evaluation of features in 18 bench mark datasets. Two approaches for selecting the optimal feature subset was used based on Grey Wolf Optimization. Experimental results were compared with Particle Swarm Optimization and Genetic Algorithm. In the proposed method, the mean fitness function for breast cancer was 0.027 and for lymphography the mean fitness achieved was 0.151.

K. B. Nahato, K. H. Nehemiah, and A. Kannan (2016) have combined the advantages of fuzzy sets and extreme learning machine to predict heart disease and diabetes. Trapezoidal member function was used to transform the clinical dataset into fuzzy sets. First five values of the nearst neighbors were used to impute the missing values in the dataset. The classification algorithm used in this work is feed forward neural network (FFNN). In this research work, the FNN was used with one hidden layer. The weights between the hidden layer neurons and output layer neurons are found based on FFNN combined with extreme Learning Machine (ELM) was used to find the weights between the hidden and output layers of the neural network. Experimental results reveal that FFNN combined with ELM was able to achieve high accuracy for Statlog Heart Disease (SHD).

SirageZeynu and Shruti Patil (2018) have proposed an ensemble model that uses Information gain attribute evaluator and wrapper subset evaluator to predict Chronic Kidney Disease. InfoGainAttributeEval with ranker and Wrapper Subset Evaluator with Best first search was used for selecting the important features. Experimental results of this research work reveal that Wrapper Subset Evaluator with Best first search engine feature selection method evaluated against K-nearest neighbor classification algorithm has shown an accuracy of 99%.

Bashir, S., Qamar, U., Khan, F. H., & Naseem, L (2016) have proposed a model based on ensemble classification which employs majority Vote Based (MV5) methodology for predicting heart disease. Five different classifiers were used in this research work to build the ensemble model. Results show that an accuracy of 88.5% with sensitivity and specificity measures of 86.96% and 90.83% respectively was achieved by the MV5 framework. The proposed ensemble framework is found to achieve better prediction of the disease after applying the model to all the selected datasets.

Elhoseny M., K. Shankar and J. Uthayakumar (2019) have proposed a method that combines Ant Colony Optimization (ACO) and density based feature selection to experiment the performance of the method for the CKD dataset. The authors were able to derive a feature set of 14 features that are significant for diagnosing CKD.

Hoque, N., Singh, M. & Bhattacharyya D.K. (2018) have proposed an EFS-MI technique for finding optimal features in Chronic Kidney Disease dataset. The authors have derived the optimal feature subset by using feature-class and feature-feature mutual information methods. The feature subset was evaluated using Decision tree, SVM, random forests and KNN.

3. METHODOLOGY

The framework of the proposed feature selection approach D-ITLBO algorithm is illustrated in Figure 1. The proposed work is carried out in three phases: Data preprocessing, Feature Selection, and Classification.

3.1 DATA PREPROCESSING

One of the major step in any information extraction process is that of data preprocessing [7]. Data preprocessing is applied as there are missing values in the datasets. K-Nearest Neighbor (KNN) methodology is applied to fill the missing values in the datasets. The K nearest values replaces the missing values in the dataset by taking the average of k nearest values [8]. The KNN method is implemented on all the five medical datasets to fill the missing values and prepare that data for the next step i.e., for feature selection.

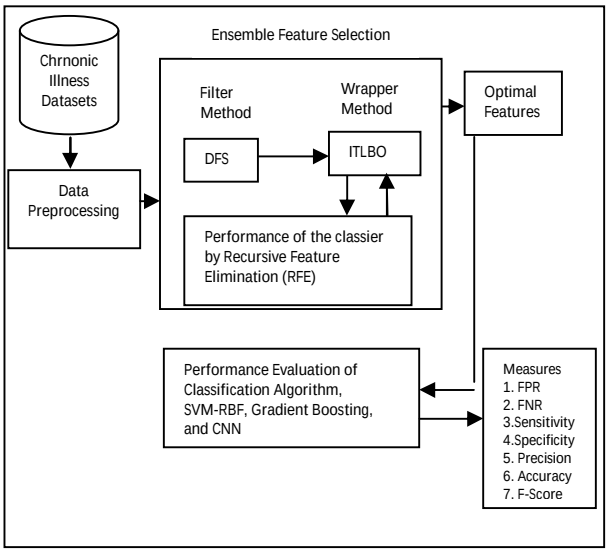

Figure 1: An Overview of D-ITLBO Method

3.2 Feature Selection

The optimal feature subset is derived with the proposed algorithm by the following processes.

3.2.1 Density Based Feature Selection (DFS) method

In this method, the merit of each feature is estimated using all classes along with their correlations. There are two popular methods for estimating the PDF of each feature in the dataset. They are the parametric and non-parametric methods [9]. In this research work, non-parametric approaches are adopted for the feature estimation and the general form of estimating PDF based on the nonparametric is given in the following equation.

$$PDF(x) \cong \frac{k}{N*V} \quad (1)$$

In (1), PDF(x) denotes the density value for feature x, V denotes the volume around x, N denotes the total number of instances and k denotes the number of instances in V.

.

The features that exhibit low density value has less significance in the diagnosis of the disease and features that show same values for their PDF are considered as redundant and hence those features with least rank and same density value are removed from the feature set.

Finally, the feature subset with highest rank without redundancy and less correlation are obtained from this filter approach. The feature subset derived from DFS is considered as input for the wrapper based ITLBO feature selection approach. The pseudo code for filter-based DFS method is given in Table 1.

Table 1: Density Based Feature Selection (DFS) method

Input: D= {d_1, d_2,…..d_N}// Dataset containing N number of instances
Input: X= {x_1,x_2,…..x_m} // Set of m features present in the dataset
Output: DFS^{ranked} // List of ranked features (desired features receive high density value)
for d= 1 **to** N
  for f = 1 **to** m
    Step 1. Estimate the Probability Density Function of feature x as PDF (x) using (1)
    Step 2. Analyze the density values of the features based on step 1.
    Step 3. Rank the features according to their density value in descending order.
    Step 4. Remove features from least rank and same density value.
    Step 5. Select the features above the threshold in DFS^{ranked}.
  end for
  DFS^{ranked}= Feature subset with high density value which are important for diagnosis.
end for

3.2.2 Improved Teacher Learner Based Optimization (ITLBO) Algorithm

The ITLBO algorithm is a variant of the TLBO algorithm [11] which finds its applications widely in many areas viz., chemical engineering, molecular engineering, medical informatics, clinical decision support systems, pattern recognition etc.. In the basic TLBO and its variants, the fitness function that corresponds to the optimal solution space is updated based on the Euclidean distance.

In the proposed ITLBO method, the existing fitness function is evaluated using the Chebyshev distance so that the algorithm converges in less time and the optimal solution space consists of less correlation between the features. Less correlation leads to reduced redundancy and hence the optimal feature subset derived for the best fitness function consists of the most discriminatory features that are non-redundant and are represented in lower dimensional space.

The formula for calculating the Chebyshev distance is given in (4). In (4), the two values x_i and x_j are used in the column vector for updating the mean values.

$$Difference1_s = r(M\_new1,s - T_F M1_s) \quad (3)$$

$$Dis_{chebyshev}(x_i, x_j) = \max(|x_i - x_j|) \quad (4)$$

$$X1'new = f1(x) + Dis_{chebyshev}(x_i, x_j) \quad (5)$$

In (3), M1_new, represents the new mean value. The value obtained from (4) used to update the previos value and the new objective function X1'new is calculated using (5). The derived objective function is taken if the value of the function is good and then the algorithm is moved to the learner phase. The algorithm comes to an end when features in the dataset that are undertaken for evaluation are updated in the solution space and the number of iterations is finally reached. The solution finally obtained forms the optimal solution of the problem.

In the proposed D-ITLBO algorithm, the results of the filter approach is provided to the ITLBO method. The wrapper method finds the optimal feature subset that are most discriminatory in diagnosis of chronic diseases. Empirical results of the proposed method are examined using SVM-RBF, Gradient Boosting, and CNN classifiers.

The experimental results of the feature selection process and the performance of the classifiers are discussed in the following section.

4. EMPIRICAL RESULTS AND DISCUSSION

4.1 Dataset Description

To investigate the performance of the ensemble D-ITLBO method, Chronic Kidney Disease (CKD), Wisconsin Breast Cancer Dataset (WBCD), Parkinsons, Thyroid Disease, and Pima-Indian diabetes dataset are taken from UCI machine learning repository are considered.

These datasets consists of numerical, categorical, Boolean type of data with varying number of instances and features. All the datasets are binary class datasets which indicate that any instance in the dataset falls into one the two classes, class '1' presence of chronic disease and class '0' not having the chronic disease. Table 2 depicts a summary of these datasets. Experiments were carried out using MATLAB tool, 2013.

Table 2: A summary of the datasets

S. No.	Dataset	No. of attributes	No. of Instances	No. of Class	Data Type
1	Chronic Kidney Disease	24	400	2	Numerical & Categorical
2	Wisconsin Breast Cancer Dataset	31	569	2	Numerical & Categorical
3	Parkinson Disease	22	196	2	Numerical & Categorical
4	Thyroid Disease	27	1000	2	Boolean, Numerical & Categorical
5	Pima Indian Diabetes Dataset	8	768	2	Numeric

4.2 Result Analysis of DFS Method

Based on (1), the density values of the features are derived. The DFS method ranks the features from those which exhibit high density value to the features that show low density value in the decreasing order. The features that have similar density values are considered to be redundant features and they are eliminated from the original feature set. The resultant features is the feature subset derived from DFS method. These derived features are then fed into the ITLBO method for further analysis. The number of features derived from the DFS method for the five datasets is depicted in Table 3.

From Table 3, it is clear that the features that are selected based on their weightage derived using equation (1) is more important and therefore they are ranked from high density value to low density value. The CKD dataset has derived a set of 19 features from 24 features. The WBCD has derived a set of 22 features compared to 31 features in the original dataset. Similarly, Parkinson's disease dataset, Thyroid dataset, and Pima-Indian diabetes dataset have derived a set of 18, 20 and 5 features respectively from the DFS method. These features are given as input to the wrapper method in the next stage for further analysis.

Table 3: Features selected based on DFS method

S. No.	Name of the Dataset	features selected	Features kist based on the ranking of the features in decreasing order of their density values
1	Chronic Kidney Disease	19	12,11,4,23,3,18,20,16,5,22,,17,2,19,7,9,21,24,1,10
2	Wisconsin Breast Cancer Dataset	22	12,13,11,27,28,26,8,7,29,6,25,9,5,30,18,17,16,10,14, 19,31,22
3	Parkinson Disease	18	10,21,22,17,20,23,9,14,11,12,13,5,7,4,15,8,6,19
4	Thyroid Disease	20	26,22,20,24,18,10,14,3,5,21,15,7,4,23,16,1,2,9,19,27
5	Pima Indian Diabetes Dataset	5	2,8,6,1,7

4.3 Result Analysis of D-ITLBO Method

The results obtained from the DFS method have to be fine-tuned to obtain an optimal feature subset that could contribute greatly to higher percentage of feature reduction

and enhanced classification performance. The features that are obtained in DFS method are evaluated using the ITLBO method and performance of the classifiers - SVM-RBF, Gradient Boosting, and CNN based on Recursive Feature Elimination (RFE) technique. Using RFE technique, the features are removed one by one from the derived feature subset. The feature subset that gives the highest classification accuracy is regarded as the best feature subset for that dataset.

Table 4 gives the feature reduction in percentage for each dataset and for each classification algorithm. The performance of the classification algorithms measeured in terms of Sensitivity, Specificity, and Accuracy are also given. The graph depicted in Figure 2 depicts the accuracy of the three classification algorithms for the five medical datasets taken for evaluation.

Table 4: Empirical results of D-ITLBO method for all the five medical datasets

S. No	Dataset	No. of Features selected using D-ITBO method	Feature Reduction (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
1.	Chronic Kidney Disease (CKD)	SVM-RBF – 8 Gradient Boosting – 9 CNN - 8	66.66% 62.50% 66.66%	93 97 97.75	92.40 96.80 97.20	94 97.33 98.67
2	Wisconsin Breast Cancer Dataset (WBCD)	SVM-RBF – 14 Gradient Boosting - 16 CNN - 14	54.83% 48.38% 54.83%	94.25 97.45 99	93.40 94.34 97.17	92.44 96.92 98.32
3	Parkinson Disease	SVM-RBF - 11 Gradient Boosting - 11 CNN - 10	50% 50% 45.45%	93 98.44 98.50	95.83 100 97.92	91.84 95.92 98.64
4	Thyroid Disease	SVM-RBF - 15 Gradient Boosting – 15 CNN -13	44.44% 44.44% 51.85%	98.96 99.50 99.90	97.60 99 99.33	98 99.25 98
5	Pima – Indian Diabetes	SVM-RBF - 4 Gradient Boosting - 3 CNN - 3	50% 62.50% 62.50%	95.05 96.25 99.72	96 96 99	93.28 96.27 99.25

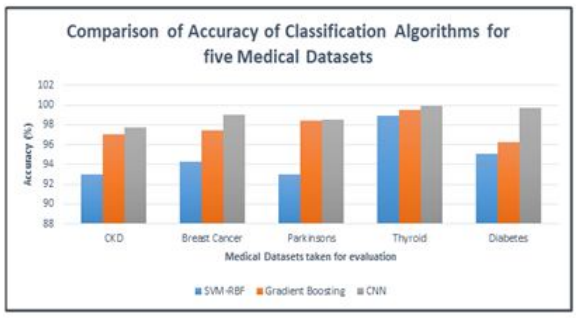

Figure 2: Graph depicting the Accuracy of the classifiers for five medical datasets

From Figure 2, it is clear that the CNN classifier has achieved hgh accuracy level compared to SVM-RBF and Gradient Boosting classification algorithms. The ensemble feature selection method – DITLBO has selected the optimal set of features with enhanced performance accuracy of the three classification algorithms. Empirical results show that the proposed ensemble feature selection technique has indeed selected few features as the required features for prediction of chronic diseases.

5. CONCLUSION

This research work predicts chronic diseases by deriving the important features significant for diagnosis of diseases. The experimental results reveal that the proposed research work reduces the original feature set to an optimal feature set that also increases the classifiers' performance. The results indicate that diagnosis of chronic diseases can be done accurately with the derived optimal feature subset and the proposed method can be tried for other chronic disease datasets also.

The proposed feature selection methodology can be further extended to compare the derived feature selection results ontology based feature selection so that the features in the datasets can be semantically analyzed. Semantic analysis of the features in the dataset will enable the developed system to handle heterogeneous data from IoT and sensor devices which forms the future scope of this research work.

REFERENCES

1. Divya Jain & Vijendra Singh. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal,19, 179-189.
2. Nilda N. dela Cruz and Ricardo Q. Camungao. (2020). Decision Support System for Predicting Cardiovascular Diseases Using Naïve Bayesian Algorithm. IJATCSE, 9(3), 3178-3183.
3. Masoudi-Sobhanzadeh, Y., Motieghader, H. & Masoudi-Nejad, A. FeatureSelect: a software for feature selection based on machine learning approaches. *BMC Bioinformatics* **20,** 170 (2019)
4. Yun Li, Tao Li & Huan Liu. (2017). Recent advances in feature selection and its applications. Knowledge and Information Systems, 53(3), 551-577.
5. Abeel T, Helleputte T, de Peer YV, Dupont P, Saeys Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26, 392–398.
6. Zhang, X., & Jonassen, I. (2018). EFSIS: Ensemble Feature Selection Integrating Stability. Computing Methodologies. 20 pages.
7. Sireesha Moturi, Srikanth Vemuru and Dr. S.N. Tirumala Rao. (2020). Classification Model for Prediction of Heart Disease using Correlation Coefficient Technique. IJATCSE, 9(2), 2116-2123.
8. Alibeigi, Mina & Hashemi, Sattar &Hamzeh, Ali. (2012). DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets. Data & Knowledge Engineering,81–82, 67–103.
9. Jamal Salahaldeen Majeed Alneamy et al. (2019). "Utilizing hybrid functional fuzzy wavelet neural networks with a teaching learning based optimization

algorithm for medical disease diagnosis." Computer Aided Design, 43 (8): 948-956.

10. J.S. Majeed Alneamy, Z. A. Hameed Alnaish, S.Z. Mohd Hashim, R.A. Hamed Alnaish. (2019). Utilizing hybrid functional fuzzy wavelet neural networks with a teaching learning-based optimization algorithm for medical disease diagnosis. Computers in Biology and Medicine, Volume 112, 2019, 103348.
11. V.R. Elgin Christo et al. (2019). Correlation-Based Ensemble Feature Selection Using Bioinspired Algorithms and Classification Using Backpropogation Neural Network. Computational and Mathematical Methods in Medicine, Volume 2019, 17 pages.
12. E. Emary, H. M. Zawbaa, and A. E. Hassanien. (2016). Binary greywolf optimization approaches for feature selection. Neurocomputing,172, 371–381.
13. K. B. Nahato, K. H. Nehemiah, and A. Kannan. (2016). Hybridapproach using fuzzy sets and extreme learning machine forclassifying clinical datasets. Informatics in MedicineUnlocked, 2, 1–11.
14. SirageZeynu and Shruti Patil. (2018). Prediction of Chronic Kidney Using Feature Selection and Ensemble Method. International Journal of Pure and Applied Mathematics, 118:24, 16 pages.
15. Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: a medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.
16. Elhoseny, M. et al. (2019). "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease." Scientific reports, 9(1): 9583.
17. Hoque, N., Singh, M. & Bhattacharyya D.K. (2018). "EFS-MI: an ensemble feature selection method for classification." Complex & Intelligent Systems, 4 (2): 105-118.