A Hybrid Multi-Level Feature Selection Framework for prediction of Chronic Disease

[1]G.S. Raghavendra , [2]Shanthi Mahesh, Professor,[3]M.V.P. Chandrasekhara Rao

RVR & JC College of Engineering, Atria Institute of technology

Abstract:

Chronic illnesses are among the most common serious problems affecting human health. Early diagnosis of chronic diseases can assist to avoid or mitigate their consequences, potentially decreasing mortality rates. Using machine learning algorithms to identify risk factors is an exciting strategy. The issue with existing feature selection approaches is that each method provides a distinct set of properties that affect model correctness, and present methods cannot perform well on huge multidimensional datasets. We would like to introduce a novel model that contains a feature selection approach that selects optimal characteristics from big multidimensional data sets to provide reliable predictions of chronic illnesses without sacrificing data uniqueness.[1] To ensure the success of our proposed model, we employed balanced classes by employing hybrid balanced class sampling methods on the original dataset, as well as methods for data pre-processing and data transformation, to provide credible data for the training model. We ran and assessed our model on datasets with binary and multivalued classifications. We have used multiple datasets (Parkinson, arrythmia, breast cancer, kidney, diabetes). Suitable features are selected by using the Hybrid feature model consists of Lassocv, decision tree, random forest, gradient boosting,Ada-boost, stochastic gradient descent and done voting of attributes which are common output from these methods.Accuracy of original dataset before applying framework is recorded and evaluated against reduced data set of attributes accuracy. The results are shown separately to provide comparisons. Based on the result analysis, we can conclude that our proposed model produced the highest accuracy on multi valued class datasets than on binary class attributes.[1]

Keywords: *Class Balance, Multi-Dimensionality, Hybrid Feature selection, Chronic diseases, Lasso CV, Ada boost, Stochastic Gradient descent,*

1. Introduction

Chronic diseases have been regarded as the most severe and lethal disease in humans. The increased rate of chronic diseases with a high mortality rate is causing significant risk and burden to the healthcare systems worldwide. Chronic diseases are more seen in men than in women particularly in middle or old age although there are also children with similar health issues When educated on appropriate data, machine learning algorithms can be excellent in identifying illnesses.[1]Heart disease datasets are freely available for model comparison. The emergence of machine learning and artificial intelligence allows academics to create the best prediction models utilizing the enormous databases that are accessible. Recent research on heart-related concerns in adults and children has stressed the need to lower chronic illness mortality. Because the available clinical datasets are inconsistent and duplicated, adequate pre-processing is critical. It is critical to choose the key traits that may be employed as risk factors in prediction models. Various supervised models using feature selection methods such as AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), Lasso Regression (LassoCV), and Random Forest (RF) are used in this work, along with classifiers. The findings are compared to previous research.[1]

2. Research Aim and Scope of the Paper

The goal of this study is to create an efficient multi-level feature selection approach that eliminates unnecessary aspects without affecting the originality of the data in order to obtain correct features that would aid in faster processing and output. The necessary steps are as follows:

1. Five datasets are used to develop a more reliable feature selection model to evaluate its performance.
2. Six selection techniques are embedded, and a hybrid model is developed to extract the most relevant features based on rank values in medical references.

Manuscript received December 5, 2023

Manuscript revised December 20, 2023

https://doi.org/10.22937/IJCSNS.2023.23.12.6

3. The performance of the framework is evaluated on binary and multi valued class datasets.

3. Literature Review

[3] Dynamic feature applications introduce additional challenges in the selection of streaming features. The dynamic features applications have several characteristics, including a) features are processed sequentially with a set number of occurrences; and b) the feature space does not exist in advance. In a text classification assignment for spam detection, for example, new features (e.g., words) are dynamically created and must thus be mined to filter out the spam rather than waiting for all features to be gathered. Traditional feature selection methods, which were not developed for streaming feature applications, cannot be employed in this setting since they require the whole feature space to be known in advance to statistically decide the significant features.[4][2] Parkinson's disease (PD) is a well-known neuro-degenerative condition. Speech/voice dysfunction is one of the early signs. Acoustic and speech signal processing technologies can evaluate and measure PD-related voice dysfunction. The current study provided a unique feature selection framework based on two stages of the feature selection approach for detecting voice loss in Parkinson's disease patients. The principle component analysis (PCA) and eigenvector centrality feature selection (ECFS) techniques are originally computed separately at the first level of selection, and the selected features from each method are treated as a distinct sub list, namely the ECFS selected features sub list and the PCA selected features sub list, in the first set.[5] [2]Traditional feature selection approaches presume that all data instances and features are known prior to learning. In many real-world applications, however, we are more likely to encounter data streams, feature streams, or both. Feature streams are defined as features that arrive one by one over time, but the number of training samples stays constant. Existing streaming feature selection approaches focus on deleting unnecessary and redundant features and picking the most relevant features, but they disregard feature interaction. A trait may have low association with the goal idea on its own, but when paired with additional features, it might be significantly connected with the target concept.[6] Feature selection is an important issue in effective machine learning, and it also adds significantly to the explainability of machine-driven judgments. During training, methods like decision trees and the Least Absolute Shrinkage and Selection Operator (LASSO) can be used to choose features. These embedded methods, however, are limited to a narrow subset of machine learning models. Wrapper-based approaches may choose features independently of machine learning models, although they are frequently computationally expensive. Many stochastic algorithms are being developed to improve their efficiency.[7] Because rolling bearings are one of the most crucial components in rotating equipment, a feature selection and fusion approach based on weighted multidimensional feature fusion is presented. To begin, features from many domains are extracted to form the original high-dimensional feature set. Given the huge number of incorrect and redundant features in the original feature set, a feature selection technique that incorporates support vector machine (SVM) single feature assessment, correlation analysis, and principal component analysis was developed. -evaluation of weighted loads.[8] multi-label feature selection has steadily gained prominence in machine learning, statistical computing, and related areas, and has been widely used for a broad range of issues ranging from music identification to text mining, picture annotation, and so on. Traditional multi-label feature selection approaches, on the other hand, use a cumulative summation strategy to construct solutions, which has the drawback of overestimating the redundancy of certain candidate features.

4. Research Methodology

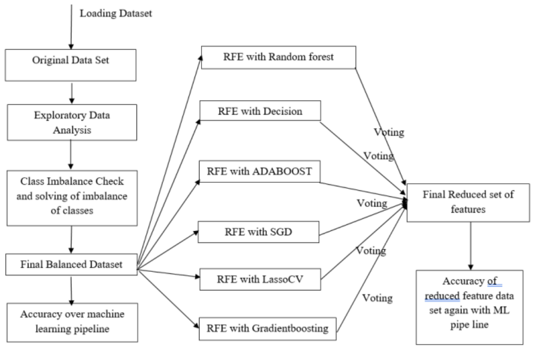
Loading Dataset
Original Data Set
Exploratory Data Analysis
Class Imbalance Check and solving of imbalance of classes
Final Balanced Dataset
Accuracy over machine learning pipeline
RFE with Random forest
RFE with Decision
RFE with ADABOOST
RFE with SGD
RFE with LassoCV
RFE with Gradientboosting
Voting
Voting
Voting
Voting
Voting
Final Reduced set of features
Accuracy of reduced feature data set again with ML pipe line

Fig 1: Working Model of Framework

Fig. 1 illustrates the working of recommended model. During data After data pre-processing and exploratory data analysis, the data set is analysed to check for missing values and negative values, which are then dealt with by the machine learning pipeline. To overcome this imbalance, to resolve these issues and avoid long execution times, three different balancing techniques are used, like smote, random, and smotek sampling techniques. This assists in making the best data set. The performance of classifiers with balanced classes will be increased. All ensemble feature selection models with classifiers are implemented to compare the binary and multi-valued class label datasets. Different training models have been given for testing the data set so that we can pick the best features for our reliable data set. Furthermore, the most appropriate features of a patient affected by chronic disease have been suggested in this diagnosis system.

5. Justification of the proposed technique

This system was built using recursive feature selection and the six ensemble feature selection methods. Numerous studies on various types of feature selection algorithms based on classifiers have already been conducted. We chose three of the most common techniques (DT, RF, and LASSO) and three less common techniques (SGD, GB, and ADA). Previous research has found that the anticipated accuracy of existing feature algorithms is high when compared to other existing methodologies. Furthermore, a small number of experiments have shown that ensemble work of wrapper-based feature selection may do rather well with very high accuracy. Except for DT and kNN, none of the study endeavours significantly approached our offered approaches as a base classier. As a result, all of the preceding approaches have been investigated further in this study using ensemble techniques to make the suggested model more efficient. Although it can be shown from the Literature Review that hypotheses put up in and yielded promising prediction accuracy, it was not high enough in contrast to our study.

6. Implementation

Simple libraries such as Pandas, Pyplot , and Scikit-learn are used to create the model, which is built in the Pyspark (Python) programming language and runs on an Apache Spark cluster.

Dataset

The first and most basic part of employing machine learning algorithms to obtain reliable results is data. The dataset used in this study was obtained from a well-known data repository, the 'UCI machine learning repository. There are six separate datasets available: Parkinson's disease, arrhythmia, breast cancer, renal disease, and diabetes... To acquire more precise results, we merged all of them in this study. More than 1190 examples from their database are compiled as a text label, together with 14 distinctive characteristics. For example, diabetes dataset’s 13 properties are used as diagnostic inputs, while the 'num' attribute is used as an output. All or most of the following six medically important variables were present in all or most records: age in years (age), sex (sex), resting blood pressure (trestbps), fasting blood sugar (fbs), chest pain type (cp), and resting electrocardiographic data (restecg). Table 1 explains the various properties and their possible values.

Data preprocessing and cleaning techniques:

In the current world, there is a significant quantity of collected data that may be obtained through the internet, questionnaires, and trials, among other means. However, the data to be used frequently contains missing values, noise, and distortions. The pooled dataset utilized for this study comprises missing or null values as well. To cope with missing values, common approaches such as imputation and deletion can be utilized.

Hybrid Feature Selection Techniques into single framework:

1) **Recursive Feature Elimination**, or RFE for short, is a prominent feature selection approach is popular because it is simple to set up and use, and it is successful in selecting those features (columns) in a training dataset that are significant in predicting the target variable. When utilizing RFE, there are two crucial configuration options: the number of features to choose and the method used to

assist in choosing features. Both hyperparameters can be investigated, albeit the method's success is not heavily reliant on them.

2) **RFE with Random Forest:** Random Forest is a well-known supervised learning machine learning method. It may be utilized in ML for both classification and regression tasks. It is built on the notion of ensemble learning, which is the process of merging numerous classifiers to solve a complicated issue and enhance the model's performance.

3) **RFE with Decision Tree:** A decision tree is a non-parametric supervised learning technique that may be used for classification and regression applications. It features a tree structure with a root node, branches, internal nodes, and leaf nodes. A decision tree starts with a root node that has no incoming branches. Outgoing branches from the root node feed into internal nodes, also known as decision nodes. Based on the given attributes, both node types undertake evaluations to generate homogeneous subsets, which are designated as leaf nodes or terminal nodes.

4) **RFE with Lasso**: “LASSO" stands for "Least Absolute Shrinkage and Selection Operator." Shrinkage is used in this model. Shrinkage essentially implies that the data points are recalibrated by applying a penalty to shrink the coefficients to zero if they are not significant. It employs the L1 regularization penalty approach. This form of regression is well-suited for models with high degrees of multicollinearity or when we need to automate specific elements of model selection, such as parameter removal or feature selection.

Lasso loss functions can be represented mathematically as:

$$|\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{m} |\beta_j|$$

Where, $\sum_{i=1}^{n} (y_i - \sum_j x_{ij}\beta_j)^2$ is Residual sum of square, $\sum_{j=1}^{m} |\beta_j|$ is sum of absolute value

5) **RFE with Adaboost:** All the trees or models in Adaboost do not have equal weights, which implies that some of the models will have greater weightage in the final model and some of the individual models will have less weightage in the final model.

6) **RFE with Gradient boosting:** Gradient Boosting is a prominent technique for boosting. In gradient boosting, each prediction corrects the inaccuracy of its predecessor. Unlike Adaboost, the weights of the training instances are not changed; instead, each predictor is trained using the predecessor's residual mistakes as labels. Gradient Boosted Trees is a method whose basic learner is CART (Classification and Regression Trees).

7) **RFE with Stochastic Gradient Descent (SGD) :**Gradient Descent is a prominent optimization strategy in machine learning and deep learning, and it may be used with most, if not all, learning algorithms. A gradient is the slope of a function. It quantifies the degree to which one variable changes in response to changes in another. Gradient Descent is a convex function whose output is the partial derivative of a set of parameters of its inputs. The steeper the slope, the higher the gradient. Gradient Descent is done iteratively from an initial value to discover the optimal values of the parameters to obtain the smallest feasible value of the given cost function For each iteration of stochastic gradient descent, a few samples are chosen at random rather than the whole data set.

7. Results and Discussion

The hybrid framework is evaluated on binary and multi-valued class attributes. It is found that the number of features of the Parkinson data set is reduced from 755 to 32 attributes. The 279 attributes in the Arrythmia data set, which is multivalued, are reduced to 16 features. 53 attributes in kidney are reduced to 6. 14 attributes in diabetes are reduced to 6. Breast cancer GSE reduced from 3000 to 158. This framework performed well on multi-value classes rather than binary class attributes.

Dataset Name	**No of original Attributes**	**Reduced attributes (Using Hybrid Feature Selection)**
Parkinson Dataset	755	32
Arrythmia Dataset	280	16
Kidney	53	6
Diabetes	14	5
Breast Cancer(GSE)	3000	158

Comparing framework with Individual methods:

Feature selection Method	**Dataset Name**	**No of attributes**	**Reduced attributes**	**Framework**
Univariate selection	Parkinson Dataset	755	98	32
RFE	Parkinson Dataset	755	128	
RFE With DT	Parkinson Dataset	755	377	
RFE With RF	Parkinson Dataset	755	245	
Univariate selection	Arrythmia Dataset	279	46	16
RFE	Arrythmia Dataset	279	78	
RFE With DT	Arrythmia Dataset	279	92	
RFE With RF	Arrythmia Dataset	279	67	
Univariate selection	Kidney	53	23	6
RFE	Kidney	53	9	
RFE With DT	Kidney	53	14	
RFE With RF	Kidney	53	19	

Conclusion

To assure the success of our proposed model, we used balanced classes by applying hybrid balanced class sampling methods on the original dataset, as well as methods for data preprocessing and data transformation, to offer credible data for the training model. We ran and evaluated our model on datasets with binary and multivalued classifications. We utilized a variety of datasets (Parkinson, arrythmia, breast cancer, kidney, diabetes). The Hybrid feature model, which includes LassoCV, decision tree, random forest, gradient boosting, Ada-boost, stochastic gradient descent, and done voting of attributes, is used to pick suitable features. The accuracy of the original dataset before applying the framework is recorded and compared to the accuracy of the reduced data set of characteristics.

References

[1] Pronab Ghosh1, Sami Azam 2, Mirjam Jonkman2, (Member, Ieee),Asif Karim 2, F. M. Javed Mehedi Shamrat3, Eva Ignatious 2,Shahana Shultana1, Abhijith Reddy Beeravolu2,And Friso De Boer2 , ”Efficient Prediction of Cardiovascular DiseaseUsing Machine Learning Algorithms With Reliefand LASSO Feature Selection Techniques”,IEEE Access , Machine Learning .

[2] Helen C. S. C. Lima 1, Fernando E. B. Otero 2, Luiz H. C. Merschmann 3,And Marcone J. F. Souza,”A novel hybrid feature selectionalgorithm for hierarchical classification”,10.1109/ACCESS.2021.3112396, IEEE Access

[3] Naif Almusallam; Zahir Tari; Jeffrey Chan; Adil Fahad; AbdulatifAlabdulatif; Mohammed Al-Naeem ,“Towards an Unsupervised Feature Selection Method for Effective Dynamic Features”, IEEE Access (Volume: 9)

[4] Amira S. Ashour; Majid Kamal A. Nour; Kemal Polat; Yanhui Guo; Wafaa Alsaggaf; Amira E,”A Novel Framework of Two Successive Feature Selection Levels Using Weight-Based Procedure for Voice-Loss Detection in Parkinson’s Disease”, IEEE Access

[5] Peng Zhou; Peipei Li; Shu Zhao; Xindong Wu,”Feature Interaction for Streaming Feature Selection”, IEEE Transactions on Neural Networks and Learning Systems (Volume: 32, Issue: 10, Oct. 2021)

[6] Zigeng Wang; Xia Xiao; SanguthevarRajasekaran,”Novel and efficient randomized algorithms for feature selection”,Big Data Mining and Analytics (Volume: 3, Issue: 3, Sept. 2020)

[7] Yazhou Li; Wei Dai; Weifang Zhang,”Bearing Fault Feature Selection Method Based on Weighted Multidimensional Feature Fusion”,IEEE Access (Volume: 8)

[8] Wanfu Gao; Juncheng Hu; Yonghao Li; Ping Zhang,”Feature Redundancy Based on Interaction Information for Multi-Label Feature Selection”, IEEE Access (Volume: 8)