A SURVEY ON FEATURE SELECTION FOR CHRONIC DISEASES CLASSIFICATION SYSTEMS

[1]Anchal Jain, [2]Prof Dr. S.M. Shah
[1]ME Student, [2]Head of Department and Professor
[1]Department of Computer Engineering,
[1]L.D. College of Engineering, Gujarat Technical University, Ahmedabad, India.

Abstract: *The mortality rate due to chronic diseases is increasing day by day. Timely diagnosis at an early stage can help in taking appropriate measures for prevention and mitigation of these diseases along with better prognosis. In this paper we present a survey on various approaches of feature selection and classification techniques used for prediction and identification of these diseases. We shall discuss the importance of feature selection methodologies for improving the accuracy and performance of classification systems. Feature selection, a dimensionality reduction technique can help deal with the problem of "curse of dimensionality" as large data sets means huge number of records and even greater number of features and also enhance its computational efficiency. The results and performance yielded by various classification algorithms after application of these pre-processing techniques gives us promising results.*

IndexTerms : Data Mining, Feature Selection, Classification, Hybrid.

I. INTRODUCTION

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. It helps to mine biological data from massive datasets gathered in biology and medicine. The insights derived via Data Mining in these datasets can be used to make useful predications in healthcare.

Chronic disease is defined as a disease that persists for a long time. It is one lasting 3 months or more, by the definition of the U.S. National Centre for Health Statistics. Chronic diseases generally cannot be prevented by vaccines or cured by medication, nor do they just disappear. They can be fatal if not diagnosed at right time.

Chronic diseases tend to become more common with increasing age. The leading chronic diseases in developed countries include arthritis, cardiovascular disease such as heart attacks and stroke, cancer such as breast and colon cancer, diabetes, epilepsy and seizures etc. Hence it becomes very important to identify these as its early detection can help in taking preventive actions and effective treatment at initial stage.

Manually monitoring the patients' data can be tiresome and requires lots of time. Hence we try to use the advancement in the field of technology to automate and speed up this process. Chronic disease classification systems also known as Chronic Disease Prediction Systems are a result of one such approach and plays a pivotal role in healthcare informatics.

Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. Classification and prediction [4,5]is a data mining technique which first uses training data to develop a model and then the resulted model is applied on testing data to get results of prediction. Various classification algorithms have been applied on disease datasets for the diagnosis of chronic disease and the results have been found to be very promising. There is an utmost need to develop a novel classification technique which can expedite and simplify the process of diagnosis of chronic disease.

II. FEATURE SELECTION FOR CHRONIC DISEASE PREDICTION

In data mining, a feature is an individual measurable property or characteristic of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective prediction. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons as simplification of models to make them easier to interpret by researchers/users, shorter training times, to avoid the dimensionality, enhanced generalization by reducing overfitting.

The focus of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results [1].More features doesn't necessarily mean more discriminative information. Hence we aim to remove those redundant and irrelevant features that confuse the learning algorithms. To remove an irrelevant feature, a feature selection criterion is required which can measure the relevance of each feature with the output class/labels. From data mining point if a system uses irrelevant variables, it will use this information for new data leading to poor generalization [2].

In this paper we will look at some of the methods developed for this purpose. In [8], the variable elimination methods were broadly classified into filter, wrapper and embedded methods.

Filter methods act as pre-processing to rank the features wherein the highly ranked features are selected and applied to a predictor. A threshold value is selected above which all features are selected. They can be easily scaled to very high-dimensional datasets, perform very fast and are computationally simple. They are not dependent on any particular algorithm and have better computational complexity as compared to the wrapper methods. However these methods do not take into account the interaction with the classifier. Also in these methods each feature is measured separately and thus does not take into account the feature dependencies. Lack of feature dependencies results in the degraded performance as compared to other techniques. Relief-F, Information Gain, Gain Ratio, Chi-Squared, Minimum Redundancy-Maximum Relevance (MRMR) etc. are some of the methods under this category.

Two different filter based approaches Relief-F and Information Gain are individually used for Cardiac Arrhythmia dataset with SVM and Logistic Regression classifiers in [16]. Classification performance was compared for accuracy, sensitivity and specificity. It was found that Relief-F with SVM gave the best results.

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features. They provide better accuracy but is dependent on classifier and are computationally extensive and have higher risk of overfitting. Sequential forward selections, backward elimination Method, Randomized Hill-Climbing Best-First Search, Branch-and-Bound Method, Genetic Algorithms etc. are some of the methods under this category.

The proposed method in [14] involved the sequential floating forward selection (SFFS) algorithm applied as a wrapper FS technique to determine the best subset of features. At the classification stage, the support vector machine (SVM), owing to its good performance in solving classification problems, was deployed. The results indicated that a combined model of SVM based on the SFFS approach can yield greater accuracy than the other methods including artificial bee colony (ABC), genetic algorithm (GA) and sequential forward selection (SFS) and principal component analysis (PCA), relief and information gain (IG) .

Embedded methods [1,9,10] include variable selection as part of the training process without splitting the data into training and testing sets. These methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods. They are computationally inexpensive than wrappers, dependencies between features can be captured effectively, provide better usage of available data and faster solutions. Also are less prone to over-fitting than wrapper techniques. But they have poor generality, selection of relevant features give due consideration to the classifier and are computationally costlier than the filter. Least absolute shrinkage and selection operator (LASSO), Elastic Net, Ridge Regression, Artificial neural networks, weighted naïve Bayes etc. are some of the methods under this category.

III. HYBRID APPROACH

In recent times, it is one of the widely used approaches used by the researchers for applying feature selection technique. The method aggregates one or more approaches together to take advantage of the merits of different approaches to get optimal results. Filter and wrapper methods are complimentary to each other but their hybrid usually achieves higher accuracy compared to wrapper methods and high computational efficiency compared to filter methods.

The hybrid feature selection technique using Relief-F-PCA method is proposed by authors in [6] for k-NN classifier. The performance of the presented hybrid framework is found to be best compared five other methods – Correlation Based feature Selection (CBS), Fast Correlation Based Feature Selection (FCBF), Mutual Information Based Feature Selection (MIFS), MODTree Filtering Approach and Relief-F Feature Selection in terms of performance metrics as Accuracy, Recall, Precision and Specificity.

Authors in [7] used hybrid of Information Gain (IG), Gain Ratio (GR), Relief-F (RF) and Chi-Squared (CS). Average of the rankings of the features on each feature selection algorithm is taken to combine results. These features are then passed to Naïve Bayes, Random Forest, KNN,MLP and SVM classifiers as single classifier, and also combined Bagging, Ensemble (MLP+SVM+KNN, MLP+SVM) classifier. This hybrid improves accuracy, f-measure, specificity, sensitivity for individual classifiers and more better for ensemble but lacks in reducing running time for some classifiers.

Four filter feature selection algorithms such as Information Gain (IG), Gain Ratio (GR), Chi-Squared (CS) and Relief-F (RF) were implemented for selecting best features for SVM classifier by authors in [8].It was found that SVM classification with gain ratio feature selection utilization contributes better performance than Information Gain, Chi-Squared and Relief-F utilization for breast cancer classification in terms of accuracy, sensitivity and AUC but could have used a wrapper method as well for improving specificity.

A new hybrid feature selection algorithm is proposed, which combines the filter algorithm based on information gain and the wrapper algorithm based on Sequential Forward Floating Search (SFFS) and Decision Tree (DT) in [13]. The experiments show that the maximum ratio of the number of reduced features and the number of initial features is 92.86%. Compared with other feature selection algorithms, the maximum decline of the number of iterations is about 67.8%, and the maximum increase of the classification accuracy is about 10.5%. The results prove that the hybrid algorithm possesses higher computational efficiency and classification accuracy.

A simple but efficient hybrid feature selection method is proposed based on binary state transition algorithm (BSTA) and Relief-F, called Relief-F-BSTA in [15]. For public datasets, the proposed method improved the classification average accuracy by about 2.5% compared with the filter method. For a specific biomedical dataset AID1284, the classification accuracy significantly increased from 77.24% to 85.25% by using the proposed method. However, it has not been applied to high-dimensional or online datasets.

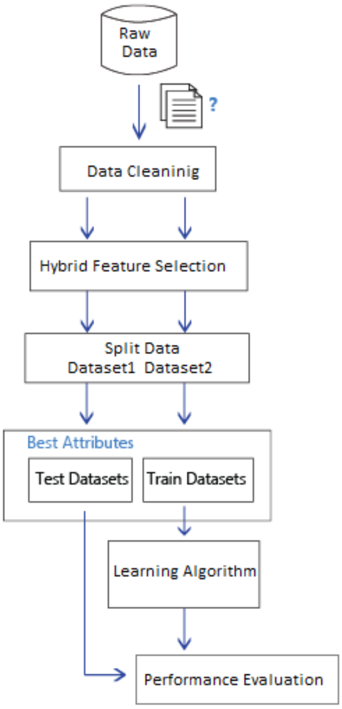

Fig 1: General Model for the Hybrid System

IV. CLASSIFICATION SYSTEMS

Healthcare industry faces many challenges in processing massive health records. Analysis of substantial amount of medical data brings complexities due to the unstructured nature of data. The health care system needs to evolve and innovate continuously to solve the problems associated with the management of huge amount of data.

In medical data mining [4–5], literature shows that many researchers used different classifier systems for chronic disease prediction to get good diagnostic results and prediction accuracy. Various classifiers like support vector machines, neural networks, decision trees, Naïve Bayes etc. [23,6] have been implemented and used in the past for the prognosis and diagnosis of chronic diseases. Fig. 2 depicts how classification process is applied on processed data to get predictive results.

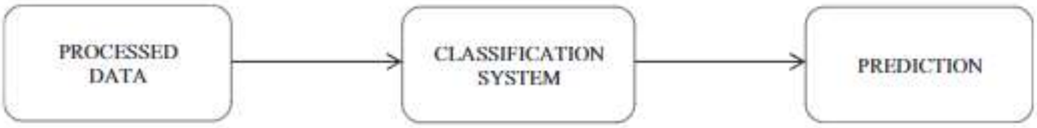

Fig 2: Classification Process

Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes and Random Forest classification algorithms are used with Minimum Redundancy Maximum Relevance Feature Selection (MRMR) feature selection in [17]. The result shows that increase in accuracy has been achieved in two of techniques Logistic Regression (SVM) and Naïve Bayes. In [18] the features were passed to Linear Discriminant Analysis to analyse the performance of the attributes by reducing it and were passed to classifiers such as Support Vector Machine, Relevance Vector Machine, Fuzzy Logic and Gaussian Naïve Bayes for detecting breast cancer and the accuracies was more than 92% with SVM being 97%(highest).

Random forest (RF) and the Generalized linear model (GLM) were used as feature selectors in [19] with classifiers K- Nearest neighbour (KNN), Neural Network (NNet), Support Vector Machine (SVM). SVM outperformed other classifier with the GLM facilitator.[20] used hybrid of Information Gain (IG), Gain Ratio (GR), Relief-F (RF) and Chi-Squared (CS) whose average rankings were used for feature selection. These features were then passed to Naïve Bayes, Random Forest, KNN,MLP and SVM classifiers as single classifier, and also combined Bagging, Ensemble (MLP+SVM+KNN, MLP+SVM) classifier. It improves accuracy, f-measure, Specificity, sensitivity for individual classifiers and more better for ensemble but lacked in reducing running time for individual classifiers.

Work has been done in [21] for the identification and classification of neuro-degenerative diseases by the use of PCA-LDA algorithm. Statistical features have been extracted, ranked, reduced using PCA and classified using LDA which improved results for ALSD-PD and ALSD-HD category. Features are selected for classification using Sequential Forward Selection (SFS) method based on 10-fold validation of the training data helps in determining an optimal suite of features for classification using linear SVM in [22].Results show that SVM is more robust and computationally faster with a large set of features and less prone to over-training when compared to traditional classifiers.

V. PERFORMANCE METRICS FOR EVALUATION

Different performance metrics are available to measure the performance of classification systems like sensitivity, precision, F-measure, accuracy and specificity, recall. These performance metrics are generally used to analyse the performance of different models.

(a)Accuracy: It is defined as the number of all correct predictions made divided by the total number of predictions made. This is defined as ratio of appropriately classified data to overall classified data.

Accuracy = (TP+TN) / (TP+FP+TN+FN)

(b)Sensitivity (Recall or True positive rate): Recall is how many relevant items are selected. It is a ratio of true positive to the sum of true positive and false negative. In medical diagnosis, test sensitivity (Recall) is the ability of a test to correctly identify those with the disease (true positive rate). If the test is highly Recall and the test result is negative you can be nearly certain that they don’t have disease.

Recall = TP / (TP+FN)

(c)Specificity (True negative rate): Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

Specificity = TN / (TN+FP)

(d)Precision: Precision is how many selected items are relevant. It is a ratio of true positive to the sum of true positive and false positive. Test specificity (Precision) is the test’s ability to correctly recognize those that do not have a disease (true negative rate). If the test output for an extremely precise test is positive user can be nearly certain that they actually have the disease.

Precision = TP / (TP+FP)

(e)False Positive Rate: False positive rate is defined as the number of incorrect negative predictions divided by the total number of negatives.

FPR = FP / (TN+FP)

(f)F-measure: The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is

$F = 1 / (\alpha\ 1/P + (1- \alpha)\ 1/R)$

where the weight α € [0, 1].

In above equations

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

VI. CONCLUSION AND FUTURE WORK

Chronic diseases have affected world badly and mortality rate due to it is increasing gradually. So it is required to develop a system that helps in their timely prediction. This article presents a survey on various feature selection and classification techniques which can be very helpful for severity analysis for quick disease diagnosis. Several reliable and efficient feature identification approaches have been developed in the literature according to different principles. Although feature selection is a well-developed field, researchers are focusing on designing novel methods to improve efficiency of the learning machines. Using hybrid mechanisms instead of individual methods gives better results especially if the methods are complementary as they make up for the demerits of each other.

In this paper we discussed about some of the different classification systems and ensembles of various classification algorithms used for predicting different types of chronic diseases and how their performance was affected due to use of individual and hybrid feature selection approaches. Overall we can conclude that feature selection do improves the performance in terms of accuracy, sensitivity and other performance metrics described in section v. Every feature selection algorithm may not enhance performance of every classification algorithm and not necessarily for each performance parameter. Future work can be done in finding the combination of feature selection and classification algorithms which can work upon all the factors at once.

VII. REFERENCES

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[2] Girish Chandrashekar, Ferat Sahin “A survey on feature selection methods”Computers & Electrical Engineering,Volume 40 Issue 1 2014

[3] Mehnaz Khan, S. M. K. Quadri ”Effects of Using Filter Based Feature Selection on the Performance of Machine Learners Using Different Datasets” BIJIT 2013

[4]Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

[5] Tan PN. Introduction to data mining. Pearson Education India; 2006.

[6] Divya Jain, Vijendra Singh “Diagnosis of Breast Cancer and Diabetes using Hybrid Feature Selection Method” 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018).

[7] Burak Kolukisa, Hilal Hacilar, Gokhan Goy, Mustafa Kus, Burcu Bakir-Gungor, Atilla Aral, Vehbi Cagri Gungor “Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease “ IEEE International Conference on Big Data (2018)

[8] T A H Tengku Mazlin, R Sallehuddin and M Y Zuriahati ”Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification” Joint Conference on Green Engineering Technology & Applied Computing 2019

[9] Langley P. Selection of relevant features in machine learning. In: AAAI fall symp relevance; 1994.

[10] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–70

[11] Jiliang Tang, Salem Alelyani and Huan Liu "Feature Selection for Classification: A Review"

[12]Divya Jain, Vijendra Singh "Feature selection and classification systems for chronic disease prediction: A review" Elsevier Egyptian Informatics Journal (2018)

[13] Jianli Ding, Liyang Fu "A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search" Journal of Intelligent Computing Volume 9 Number 3 September 2018.

[14] Saeid Fallahpour, Eisa Norouzian Lakvan, Mohammad Hendijani Zadeh"Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem" Elsevier Journal of Retailing and Consumer Services 34 (2017)

[15] Zhaoke Huang, Chunhua Yang, Xiaojun Zhou,Tingwen Huang"A hybrid feature selection method based on binary state transition algorithm and ReliefF" IEEE Journal of Biomedical and Health Informatics[2018]

[16] Alaa Elsayyad,Mujahed Al-Dhaifallah, Ahmed M. Nassef "Feature Selection for Arrhythmia Diagnosis using Relief-F algorithm and Support Vector Machines"14th International Multi-Conference on Systems, Signals & Devices (SSD) [2017]

[17]Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir "Improving Heart Disease Prediction Using Feature Selection Approaches" 16th IEEE International Bhurban Conference on Applied Sciences & Technology (2019).

[18]B.M.Gayathri, Dr.C.P.Sumathi "Feature selection using Linear Discriminant Analysis for breast cancer dataset" IEEE International Conference on Computational Intelligence and Computing Research(2018).

[19]Dr. J. Arunadevi, K. Ganeshamoorthi "Feature Selection Facilitated Classification For Breast Cancer Prediction" IEEE Xplore Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)

[20]Burak Kolukisa, Hilal Hacilar, Gokhan Goy, Mustafa Kus, Burcu Bakir-Gungor, Atilla Aral, Vehbi Cagri Gungor "Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease " IEEE International Conference on Big Data (2018)

[21]Sobia Amin, Abhishek Singhal "Identification and Classification of Neurodegenerative Diseases using Feature Selection through PCA-LD" 4th IEEE International Conference on Electrical, Computer and Electronics(2017).

[22] Barath Narayanan Narayanan, Russell C. Hardie, Temesguen M. Kebede "Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection" IEEE National Aerospace and Electronics Conference, Dayton (NAECON 2018)

[23] Pujari AK. Data mining techniques. Universities press; 2001