A STUDY ON FEATURE SELECTION FOR CHRONIC DISEASE PREDICTION

[1]M Kavitha, [2]Dr S Subbaiah
[1]Assistant Professor &Ph.D Part time Research Scholar, [2]Assistant Professor
[1]PG & Research Department of Computer Science and Applications, Vivekanandha College of Arts and Science for Women, Tiruchengodu, Tamilnadu, India
[2]Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamilnadu, India

ABSTRACT

Chronic Disease Prediction shows a fundamental role in healthcare informatics. It is vital to identify the disease at an initial stage. This paper describes a survey on the utilization of feature selection and classification techniques for the diagnosis and prediction of chronic diseases. Adequate selection of features showsanimportantpart for improvingprecision of classification systems. Dimensionality reduction supports in refining overall performance of machine learning algorithm. In this paper, we present a completesummary of various feature selection methods and their inherent pros and cons.

KEYWORDS: Feature Selection, Chronic Disease Prediction, Wrapper Method, Filter Method, Embedded Method.

1. INTRODUCTION

Diagnosis of chronic diseases is very critical in the medical field as these diseases persist for long time. The leading chronic diseases contain diabetes, stokes, cardiovascular disease, arthritis, cancer, hepatitis C. Initialfinding of chronic disease reliefs in taking precautionaryactivities and actualhandling at an earlyphase has always been found to be caring for patients. Currently, maintenance of clinical databases has developed a vital task in medical field. The patient data containing of numerous features and diagnostics related to disease should be entered with utmost attention to offerexcellencefacilities. As the data stored in medical databases may comprise missing values and redundant data, mining of the medical data becomes unwieldy. As it can disturb the outcomes of mining, it is critical to have noble data planning and data reduction earlierrelating data mining algorithms. Prediction of disease becomes rapid and easier if data is exact and reliable and free from clatter.

Feature Selection is awell-organized data preprocessing method in data mining for dipping dimensionality of data. In healthanalysis, it is very vital to classify most significant hazardissueslinked to disease. Applicable feature identification supports in the elimination of needless, joblessfeatures from the disease dataset which, in turn, gives rapid and improvedconsequences. Classification and prediction is a data mining method which major uses working out data to grow a perfect and then the causedperfect is practical on challenging data to get results of prediction.

Numerous classification algorithms have been functional on disease datasets for the diagnosis of chronic disease and the results have been establish to be very talented. There is agreatestessential to develop anoriginal classification technique which can accelerate and shorten the procedure of diagnosis of chronic disease. In this stage of data explosion, voluminous amount of medical data is formed and modernized daily. Healthcare data comprises Electronic Health Records (EHR) which includes of clinical reports of patients, analytic test reports, doctor's prescription; information related to pharmacy, information related to patient's health insurance, uprights on social media such as blogs, tweets. There is ahighestessential of awell-organized parallel data processing system which is talented to manage and examine the vastsizes of healthcare data. Chronic Disease Diagnosis (CDD) systems can be used as valuable tools for proper controller and supervision of the chronic disease. It screens the healthiness of patients and supportsurgeons

and medical careers to provide 24/7 healthcare facilities.

This paper is organized as follows. Firstly, a smallnarrative of feature selection for chronic disease prediction is presented. Secondly, several feature selection approaches and correlatedeffort on many feature selection methods is presented. Laterally with that, a study containing of countless feature selection algorithms, characteristics, facts, disadvantages.

2. FEATURE SELECTION FOR CHRONIC DISEASE PREDICTION

Feature selection, also identified as Variable Selection, is awidelyrecycled data preprocessing procedure in data mining which is principallycastoff for reduction of data by rejectingunimportant and extra attributes from any dataset. Additionally, this procedureenriches the clarity of data, simplifieswellpicturing of data, diminishes training period of learning algorithm and increases the performance of prediction. There occurplentiful applications of applicable feature identification procedures in healthcare division. Filter methods, wrapper methods, ensemble methods and embedded methods are specific of the commonly used methods used for flexible selection.

In current years, maximum of the authors are concentrating on hybrid methods used for feature selection. Beforehand any typical is applied to the data, it is always better to eliminate noisy and inconsistent data to get additional accurate results in a smaller amount time. Dipping the dimensionality of a dataset is of dominantsignificance in real-world applications. Moreover, if most significant features are selected, the complexity decreases exponentially. In new years, numerous feature identification methods have been practical on healthcare datasets to get more appreciated information.

The application of feature selection approaches is finished on clinical databases for the prediction of plentiful chronic diseases like diabetes, heart disease, strokes, hypertension, thalassemia etc. Countless learning algorithms work proficiently and give additional accurate results if the data comprisesadditional significant and non-redundant attributes. As the medical datasets coversgreat number of redundant & irrelevant features, awell-organized feature selection method is needed to excerpt interesting features relevant to the disease.

Anextremely accurate diagnostic system for the finding of knee joint disorders using VAG signals was proposed. The procedure was established using anoriginal feature selection and classification procedure. Intended for the sympathy of most significant and steady features, apriori algorithm and genetic algorithm were used. To assess their performance, random forest and LS-SVM classifiers were used. Moreover, the concept of wavelet decomposition was used to classify normal VAG signals from abnormal ones.

A comparison of the results based on evaluation metrics revealed that the performance of LS-SVM using the apriori algorithm was the greatest with an accuracy of 94.31%. The planned approach could be of excessive help for early diagnosis of knee jointcomplaints so that action can be providing to patients at an early stage. Anuncomplicatedorganization of feature selection and various gene selection methods were reviewed. Authors classified these methods under three divisions – supervised;semi supervised and unsupervised feature selection.

Various tests and difficulties in removing knowledge from gene expression data were also spoken. Certain of the plain issues raised were

(1) Dipping dimensionality of data with hundreds of thousands of features

(2) How to lever mislabeled, inexact data

(3) How to deal with extremely excessive data

(4) Determining the gene relevancy/redundancy and removing relevant biological information from the gene expressions.

It was exposed through relative study on gene selection that the classification accuracy of semi supervised and unsupervised methods were as talented as supervised feature selection. A new feature selection approach using SVM ranking with regressive search method was

presented to final the ideal subset of features on type II diabetes dataset. With the future approach, the predictive accuracy of Naïve Bayes classifier got significantly amplified. The methodology used was very humble yet actual which would certainlyassistance the physicians and medical professions for the diagnosis of Type 2 diabetes.

Modified FAST is a fast and well-organized feature identification algorithm which was proposed. Aperfect value of beginning with the inclusion of symmetric uncertainty (SU) was appropriately found. The minimum spanning tree was createdafterwardsmearing symmetric uncertainty (SU). The comparison of the results of the proposed algorithm was made with other algorithms like FAST, FCBF, Relief and CFS based on classification accuracy and percentage of features selected and it was confirmed that Modified FAST was the best algorithm among all.

3. FEATURE SELECTION APPROACHES

Traditional feature selection methods for machine learning are approximatelycategorized into three groups:

(a) Filter method

(b) Wrapper method

(c) Embedded method

Currently, hybrid methods consisting of combination of these approaches is also used by many authors the results of which are also promising.

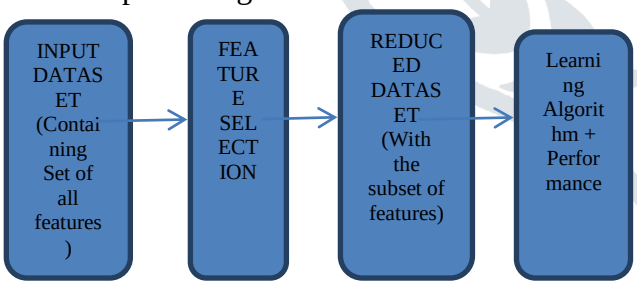

Fig. 1. Feature selection process

Fig. 1 demos the feature selection procedure that can be practical on involvement dataset to get reduced dataset which is then approved to the learning algorithm.

A. Filter Method

It is one of the eldestapproaches of feature selection. In variable selection using filter approach, filtering of features is completedbeforehand the operation of any learning algorithm. It ranks features founded on a confident evaluation criteria. As it is not reliant on on the classifier applied, it inclines to stretchmixed performance on prediction. These approachesprovide fast and efficient results on execution. So, they are favored for huge databases done wrapper methods.

The constraint of these styles is that they overlook interaction amongst classifiers and dependency of one features over additional and may fail to select the most ‘‘useful” features. MIFS (Mutual Information based Feature Selection), proposed by Battiti, is a feature selection approach created on the concept of mutual information that does ‘‘greedy” selection of features. This technique does feature identification in such a way that it excerpts maximum mutual information. However, due to the presence of great number of errors in its implementation, MIFS is less preferred. MIFS-U is a modified feature selection approach over MIFS method which was developed to make considerable use of the mutual information. The performance of MIFS-U is like an ‘‘ideal greedy selection algorithm” when there exist constant distribution of information. This approach can be efficiently used to resolve large problems. The performance of MIFS-U damages if information distributions of features deviate from the uniform distribution.

MICC (Mutual Information-based Constructive Criterion) is a greedy filter feature selection approach founded on the idea of mutual facts that was established to overwhelm the limits of its examples. The most significant feature is that it reflectssignificance as well as non-redundancy of the features to the output classes. The major advantage over its precedent algorithms MIFS and MIFS-U is that it selects features short of using any parameters such as B (Beta). So the results were more promising with MICC as compared to its examples.

Correlation based feature selection approach was realistic for the diagnosis of Coronary Artery Disease (CAD) though a hybridized model. Most important risk factors related to CAD disease were recognized using correlation feature selection approach along with particle swam optimization method followed by a clustering algorithm. In order to hypothesis diagnostic models for CAD disease, C4.5

algorithm, multi-layer perceptron (MLP), multinomial logistic regression (MLR) and fuzzy unordered rule induction algorithm (FURIA) were recycled.

The CAD model was authorized with 10-fold cross validation technique. The predictive truth of MLR algorithm was the main while it was lowest with MLP algorithm on both clinical data and Cleveland heart disease data. The consequences of the proposed methodology were very promising which significantly better the accuracy of classifier. Consequently, this method can be recycled as a valued tool for clinical decisions related to CAD disease diagnostics. Yu and Liu intended a correlation founded filter approach to deal with the problems of great dimensionality. Authors presented the concept of 'predominant correlation' for the documentation and elimination of irrelevant and redundant features and implemented fast correlation-based feature selection (FCBF) algorithm. The consequences of the testsexposed that the proposed algorithm performed with less quadratic time complexity and was very well-organized to deal with high-dimensional data.

B. Wrapper Method

Wrapper methods does selection of features by charitable due reflection to the learning algorithm to be used. The major advantage over filter methods is that it bargains the most ''useful" features and does optimal selection of structures for the learning algorithm. Furthermore, it reflectsneedsamongst features and provides more accurate results in comparison to filter methods. Conversely, it has a problematic that if another learning algorithm has to be developed, this method wants to be re-executed.

Moreover, this method is very difficult and more disposed to to over-fitting on small training datasets. A full analysis and judgment of wrapper feature selection method and relief algorithm (a filter feature selection approach) was done. Authors discovered the strengths and limitations of the wrapper approach for best feature subset selection. The trials were conducted with both real and artificial datasets beside with two induction algorithms namely, Naïve Bayes classifier and decision trees. It was exposed from results that the error rate was significantly reduced when wrapper approach was used with Naïve Bayes classifier. Maldonado et al. practicalwrapper approach using sequential backward elimination.

The techniquerecycled support vector machines and kernel functions for implementation. The future methodology presented an effective validation error measure for the removal of features. Moreover, the key aspect was that it could be used with any of the kernel functions. The significant feature of the algorithm was that each run of algorithm designateddissimilar set of features. The contrast of the results exposed that the future wrapper algorithm showedhealthier performance than current filter and wrapper methods. Though, due to backward removal of features, it was exclusive to use this approach if there were huge number of input features.

C. Embedded Method

In embedded feature selection approach, search is usually directed by the learning process. This method, also known as nested subset method, usually measures the ''usefulness" of feature subsets and performs feature selection as a part of the training process. They usually work giving to specific learning algorithm which assistances in enhancing the performance of a learning algorithm. This technique make healthier usage of existing data and offers faster solution as they do not needsplitting of training data into training set and validation set. They are computationally cheap and less prone to over-fitting than wrapper techniques. Furthermore, the computational difficulty of embedded methods is better than wrapper methods. The major constraint with these methods is that it takes conclusions depending on the classifier.

Hence, selection of features can be affected by the theory that the classifier makes which might not work with some other classifier. An embedded method based on regressive feature selection was proposed by Maldonado et al. The purpose was to select most significant features from excessive data for applying binary classification using support vector machines. The future method was very flexible and facilitated to be used with several objective functions. With the use of

fixed feature selection process, the proposed strategy reached very good results on highly imbalanced data sets. ESFS (Embedded Sequential Forward Selection), proposed by Xiao et al., is a novel embedded selection technique which was onlyfounded on incrementally adding the furthermostappropriate features.

This method was concerned with the use of mass purposes introduced from the concept of indication theory that helped the merging of information provided by features. The proposed method significantly betters the classification correctness and was able to choice the most discriminative features when practical to emotional classification (speech and music samples). With the new results, the proposed embedded method (ESFS) was found to give inferior computational cost than the wrapper method (Sequential Forward Selection).

D. Hybrid Method

In recent times, it is one of the extensively used methods used by the researchers for applying feature selection technique. The method sums one or more methods together to take advantage of the qualities of dissimilarmethods to get best results. These methods usually achieve advanced accuracy associated to wrapper methods and high computational productivity compared to filter methods. A hybrid feature selection approach founded on improved particle swarm optimization algorithm. Scholar'spractical filter and wrapper methods together for image steganalysis. It was found from the experimental results that the proposed crossed approach significantly reduced the number of features and enhanced the classification accuracy as compared to other preceding feature selection algorithms.

Also, computational cost and time also got reduced with the proposed methodology. BBHFS (Boosting Based Hybrid Feature Selection), proposed by Das is a fast and scalable hybrid algorithm which involved the idea of boosting and advantages of both filter and wrapper methods. Authors offered a more informed filter method by incorporating forward selection algorithm and certain of the benefits of wrapper method such as natural stopping criterion. This algorithm produced fast and better results than wrapper methods when applied on DNA dataset using Naive Bayes classifier and on the Chess dataset using ID3 algorithm. The approach significantly improved the performance of these classifiers. The future hybrid approach was found to be very scalable on datasets consisting of large number of features.

4. Conclusion

World's health is badly affected by the chronic diseases which isdispersal and cumulativeday by day. The absence or delay in correcttreatment can also lead to the death of patients. So, chronic diseaseprediction is a vital job in medical field. This paper presents a study on numerous feature selection and classification techniqueswhich can be very helpful for severity examination for quick disease diagnosis. Several consistent and effectivefeature identificationmethods have been developed in the literature according to dissimilarprinciples. Though feature selection is a well-developedfield, researchers are focusing on designing novel approaches toprogress efficiency of the learning machines. This study shows thatthere is a need to make healthcare professionals aware of dependablefeature selection and classification methods that can be successfullyapplied on medical databases for the early detection ofdiseases.

REFERENCES

[1] Shardlow M. An analysis of feature selection techniques.The University ofManchester; 2016.
[2] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1(3):131–56.
[3] Tang J, Alelyani S, Liu H. Feature selection for classification: a review. DataClassif: AlgorAppl 2014;37.
[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] Dunham MH. Data mining: introductory and advanced topics. PearsonEducation India; 2006.
[7] Muni Kumar N, Manjula R. Role of Big data analytics in rural health care – astep towards svasthbharath; 2014.
[8] Hussein AS, Omar WM, Li X, Ati M. Efficient chronic disease diagnosisprediction and recommendation system. In: Biomedical engineering

andsciences (IECBES), 2012 IEEE EMBS conference on. IEEE; 2012. p. 209–14.
[9] Guyon I, Elisseeff A. An introduction to variable and feature selection. J MachLearn Res 2003;3(Mar):1157–82.
[10] Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classificationmethodology for the detection of knee-joint disorders. Comput MethodsPrograms Biomed 2016;127:94–104.
[11] Ang JC, Mirzal A, Haron H, Hamed H. Supervised, unsupervised and semisupervisedfeature selection: a review on gene selection. IEEE/ACM TransComputBiolBioinform 2015;PP(99).
[12] Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. SVM rankingwith backward search for feature selection in type II diabetes databases. In:Systems, man and cybernetics, 2008. SMC 2008.IEEE international conferenceon.IEEE; 2008. p. 2628–33.
[13] Nagpal A, Gaur D. ModifiedFAST: a new optimal feature subset selectionalgorithm. J Inform Commun Convergence Eng 2015;13(2):113–22.
[14] Battiti R. Using mutual information for selecting features in supervised neuralnet learning. IEEE Trans Neural Networks 1994;5(4):537–50.
[15] Kwak N, Choi CH. Input feature selection for classification problems. IEEETrans Neural Networks 2002;13(1):143–59.
[16] Huang J, Cai Y, Xu X. A filter approach to feature selection based on mutualinformation. In: 2006 5th IEEE international conference on cognitiveinformatics, vol. 1. IEEE; 2006. p. 84–9.
[17] Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronaryartery disease cases using non-invasive clinical data. J Med Syst 2016;40(7):1–7.
[18] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlationbasedfilter solution. In: ICML, vol. 3; 2003. p. 856–63.
[19] Kumar V, Minz S. Feature selection. SmartCR 2014;4(3):211–29.
[20] Kohavi R, John GH. Wrappers for feature subset selection.ArtifIntell 1997;97(1):273–324.