Feature Selection And Dimensionality Reduction Methods For Chronic Disease Prediction

Rekha Yadav, Subhash Chandra Jat

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 the prevention and mitigation of these diseases along with better prognosis. A study on the use of selection techniques for diagnosis & prediction of chronic diseases will be submitted in this document. Adequate choice of characteristics plays an important role in increasing classification systems’ accuracy. Reduced dimensionality helps to improve machine learning algorithms overall performance. Using classification algorithms on the datasets of diseases provides promising results by designing chronic disease adaptive, automated and intelligent diagnostic systems. This work provides an overview of the various methods for selecting features and reducing dimensions.

Index Terms—Data Mining, Chronic Disease, Disease Diagnosis, Feature Selection, Feature Selection Methods, PCA, Dimensionality Reduction.

1 INTRODUCTION

THE exploration of information in databases is data mining. Data mining methods help to process the data and to make it usable. Data mining prediction outcomes are useful in various areas such as BUI, IT, healthcare management, finance, etc. Prediction results. A large number, as well as a range of processing data, are available in the medical field and there are many difficult tasks. This area needs to be handled correctly and on time. which can save many patients’ life. Data mining techniques play a vital role in healthcare analysis. Early detection and accurate results are achievable by physicians using data mining algorithms. For the diagnosis of various diseases, specific algorithms are used. Accuracy and efficiency vary depending on the data used. Medical data mining increases patient care efficiency or health trends prediction. Data mining tools help to find hidden trends, a group linked objects or determine health-care issues. For medical data, it may be necessary to apply data mining techniques like classification and estimation, clustering, association rules and various mining methods. The collected data in the form of images can also be used to mine healthcare data. Nowadays, many image mining techniques improved disease prediction and health care decision making the task as easiest. This paper discusses the use of medical data mining algorithms [1]. Machine learning (ML) is a significant method for data analysis that iteratively learns from the available data with the aid of learning algorithms. The learning process begins with comments or information, as examples, direct experience or instruction, in particular, look for data patterns and make better decisions based on the examples we provide in the future. The key goal is to allow the machines to learn automatically without human interference or assistance and adjust the response accordingly. The intended contribution of AI in the field of medical science is to develop programs that can actually help a medical expert in practicing expert and more accurate diagnosis. The forecast for diseases plays an important role in machine learning. Various types of diseases can be predicted using ML techniques. Here we examine how machine learning techniques are used to predict various disease types. This paper focused on the prediction of chronic kidney disease, heart disease, diabetes and breast cancer lymphatic system and lung disorders.[2]. In the Winnipeg Health Region and across Canada, chronic disease is a severe and growing problem. As the healthcare landscape has changed because policymakers are seeking better ways to respond to changing population needs, too much of the illness we see today is of a chronic nature. This article seeks to provide decision-makers with a better path by synthesizing scientific evidence of the best approach to managing or preventing chronic diseases [3]. By the U.S. definition, a chronic illness lasts three months or longer. Health Statistics Center Local. Typically, chronic diseases can not nor do they simply go away through vaccinations or treatment. At least one chronic health disorder (as of 1998) existed in 88% of Americans aged over 65, respectively. The leading chronic diseases contribute significantly to health-conscious activities— in particular, tobacco use, physical activity impairment and poor eating habits. The age, in general, is the origin of chronic diseases. Chronic diseases are generally defined as conditions that last for 1 or more years and require continuous medical care or limit daily living activities or both. Chronic diseases such as heart disease, cancer, or diabetes are leading causes of death or disability in the US. They are also leaders in annual healthcare costs of 3.5 trillion dollars for the country. [4].

2 CHRONIC DISEASES

Chronic diseases are long-lasting illnesses. Chronic diseases such as heart disease, cancer, chronic kidney disease, asthma, hyperlipidemic arthritis, heart disease, diabetes, prostate cancer, hemophilia, chronic respiratory disease are the most frequently seen chronic diseases. In developing countries, the most important chronic diseases include (alphabetically) arthritis, cardiovascular diseases, like cardiovascular attacks or stroke, breast and colon cancer, diabetes, epileptic or seizure diseases, obesity and oral problems. In the USA (and in other developed nations, each of these problems plagues older children) [4].

A shortlist of risk behaviors is responsible for many chronic diseases:

1. Use of tobacco or second-hand smoke exposure.
2. Poor quality, including poor fruit and vegetable diets
and high levels of sodium and fats.
3. Failure to exercise physically.
4. Too much alcohol intake.

Examples of common chronic diseases and early indications are given below.

2.1 Arthritis
In the United States, almost 43 million Americans are the leading cause of disability. Even if economic measures to reduce the burden of arthritis are effective, they are not used properly. Regular or moderate exercise gives arthritis a range of advantages by minimizing joint pain and stiffness, building strong joint muscles and enhancing strength and endurance:
1. Joint pain for a short period of time after sitting or waking up.
2. Joint rigidity.
3. Attachment tenderness and redness.
4. Joint swelling.

2.2 Cancer
The second leading cause of death in the US is cancer. The prevention, early detection or treatment of cancer is essentially controllable. For that, the national cancer burden, behavioral and environmental factors that increase the risk of cancer need to be decreased. This also calls for the provision and availability of high-quality cancer screening services or treatments, in particular in under-served populations:
1. Excessive exposure to sunlight.
2. Lack of physical activity.
3. Family history.
4. Exposure to certain chemicals.
5. Poor diet

2.3 Stroke
Blocking blood flow to the brain is a stroke. Its signs contain:
1. Sudden stomaching or weakness, especially of one part of the body of the face, arm or leg.
2. Sudden misunderstanding, the question of expression or comprehension.
3. Sudden pain in one eye or both.
4. Surgical walking problems, dizziness, balance loss or coordination.
5. Sudden serious, unexplained cause headache.

2.4 Heart attack
A blocking blood flow to the heart is a cardiac attack. Her signs of alarm could comprise:
1. Shortness of breath.
2. Irregular heartbeats.
3. Dizziness.
5. Uncomfortable pressure or pain in the chest.

2.5 Obesity
Obesity is an over-normal chronic health condition. It is a risk factor for chronic conditions such as hypertension, heart attack, heart attack, asthma, type 2, diabetes, etc. Obesity is often linked to the lifestyle of a person like:
1. Failure to exercise physically.
2. High calorie/fat diet.

2.6 Many of these chronic diseases that contribute to everyday life have rising risk factors:
1. Smoking.
2. Excessive alcohol.
3. Excessive exposure to sunlight.
4. Depression.
5. Stress. [6,7]

3 Disease Diagnosis
Disease diagnosis is abbreviated as Dx or Ds. This is the process by which the symptoms of a person can be determined. Many signs and symptoms are unspecific and therefore the diagnosis is the most difficult task. We can do the disease diagnosis using Machine Learning techniques. We can develop a model in which the user can enter his symptoms and the model gives a particular disease [8]. Diabetes mellitus is actually called diabetes, which in India is now commonly replicated but vigorously practiced as a chronic metabolic disorder. It is due to inadequate secretion or impaired insulin action. Diabetes mellitus consist of a number of conditions characterized by high blood glucose abnormality and altered carbohydrate, lipid/protein metabolism characterized by hyperglycemia, inadequate insulin resistance or insulin or both contributing to dltürar dioxide metabolism, proteins, Diabetes mellitus, Highly characterized by hyperglycemia, poor tolerance to insulin or both, causing altered carbohydrate metabolism, protein, lipids or increased risk of vascular complications, reactive oxygen species defects and high stress-mediated pancreatic beta-cell damage [9].

3.1 Algorithm in ML for Disease Diagnosis
There are several algorithms used in two phases of disease prediction that are mainly classified into two phases:

3.1.1 Principal Component Analysis (PCA)
In this method, we use orthogonal transformation which is a statistical technique for converting a possible set of correlated values known as principal components.

3.1.2 Linear Discriminant Analysis (LDA)
It is a strategy utilized as a part of insights for pattern recognition to discover the linear combination of features in a machine learning approach. The result of this will be used for linear classification. In the next Phase contains several machine learning classification algorithms as:

3.1.2.1 Decision Tree
Decision Tree is one of the most common classification algorithms included. This algorithm uses a tree building technique to divide and conquer. There are a number of events pertaining to a collection of attributes. A decision tree contains nodes or leaves in which nodes are tests on characteristic estimates or attributes, but leaves are groups of examples that satisfy conditions given. the conclusion could be ‘true’ or ‘false’ Regulations can be derived from the way that starts at the root node, ends at the leaf node as well as uses transit nodes as preconditions for the rule to predict a leaf class. cutting of trees must be carried out in order to remove unnecessary requirements and duplications.
3.1.2.2 K-Means

It can be said as a k-means grouping. This is a technique for Vector Quantization, which is initially from signal processing, which is known for cluster analysis in information mining. We can utilize the 1-closest neighbor classifier on the cluster centers that are acquired from k-means to group the new information into effectively existing groups.

3.1.2.3 KNN

It can be called as K nearest neighbor. It is the clustering technique used for the clustering of data. It can be considered as another version of K-Means. It doesn’t use the mean and distance. Instead, it is based upon voting of the nearest neighbors in the k-clusters [10].

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Units for Magentic Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support vector</td>
<td>Merits</td>
</tr>
<tr>
<td>machines</td>
<td>Gives versatility with</td>
</tr>
<tr>
<td></td>
<td>kernel concept, less</td>
</tr>
<tr>
<td></td>
<td>adaptability, rugged</td>
</tr>
<tr>
<td></td>
<td>noise, etc.</td>
</tr>
<tr>
<td>k-nearest</td>
<td>Facilitates kernel definition</td>
</tr>
<tr>
<td>neighbor</td>
<td>reduces fit, robust to noise,</td>
</tr>
<tr>
<td>algorithm</td>
<td>easy to understand and</td>
</tr>
<tr>
<td></td>
<td>enforce, strong to</td>
</tr>
<tr>
<td></td>
<td>rude training data.</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Facilitates kernel definition</td>
</tr>
<tr>
<td>classifier</td>
<td>reduces fit, robust to noise,</td>
</tr>
<tr>
<td></td>
<td>easy to understand and</td>
</tr>
<tr>
<td></td>
<td>enforce, strong to</td>
</tr>
<tr>
<td></td>
<td>rude training data.</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Provides high accuracy in</td>
</tr>
<tr>
<td></td>
<td>forecasting but</td>
</tr>
<tr>
<td></td>
<td>arbitrarily variable</td>
</tr>
<tr>
<td></td>
<td>input-output relationships</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Supports the missing data</td>
</tr>
<tr>
<td>networks</td>
<td>extensively</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>C4.5 classifier</td>
<td>Works well both</td>
</tr>
<tr>
<td></td>
<td>categorically or</td>
</tr>
<tr>
<td></td>
<td>continuously and</td>
</tr>
<tr>
<td></td>
<td>makes bigger execution of the</td>
</tr>
<tr>
<td></td>
<td>program less memory</td>
</tr>
</tbody>
</table>

4.1 Objectives of Principal Component Analysis

1. PCA reduces space for attributes from the larger number of variables to a smaller number of factors, and as a result, is a "non-dependent" process.
2. PCA is a method for reducing dimensionality or for compressing data. The aim is to reduce the size and there is no guarantee that the size is interpretable (a fact that (amateur) statisticians often do not understand).
3. Take a subset of variables from a wider sequence, that indicates the highest correlations between the original and the primary component.

4.2 Properties of Principal Component

Theoretically, a significant component can be represented as a linear combination with an optimum weighting of variables observed. The PCA output is the most important component, and the original variables are less or similar. Less, when the dimensions in our dataset are to be discarded or reduced. The PCs have some useful features described below:

1. PCs are simply a linear combination of the initial variables. In this combination, the weight vector is basically the own vector that satisfies the lesser-square definition at the same time.
2. As already mentioned, the PCs are orthogonal.
3. The variability on PCs decreases as we pass from the first to the last PC, which is why this is so significant. Sometimes the least important PCs can be used in regression, outer detection, etc. [12].

5 Feature Selection

The collection of apps, one of the most renowned approaches used to learn computers, helps to reduce the data dimension. It chose suitable features and removes obsolete features. It helps to reduce the attributes that negligible affect the health of the patient in medical diagnosis. Several researchers have studied and published productive findings in the field of selection of features in the past. Discussed the importance of the selection of features and high dimensionality problems. Researchers offered a summary of concepts, problems and real-life feature selection applications. Selection of features and extraction methods using microarray datasets have been studied and compared. Authors presented a comprehensive survey to minimize dimensionality explaining how classification accuracy can be improved and code complexity reduced. Feature selection (called subset selection) is a method widely used in machine learning, in which the features of the data are chosen for an algorithm of learning. The best subset contains the smallest number of dimensions which contribute most to the precision; the remaining, unimportant dimensions are discarded. This is one of the two ways of avoiding a dimension curse (the other is the extraction of the feature) and is an important period of preprocessing. Two methods are available:

1. Forward selection: Begin without any variables then add them one by one, adding one that decreases errors the most at each point, until any additional error doesn’t decrease significantly.
2. Backward selection: Start by removing all the variables one by one, removing one which reduces the error the most (or only slightly increases it), until the additional removal significantly increases the error [13].
5.1 Feature Selection for Chronic Disease Prediction
The feature selection is a commonly popular pre-processing technique, also called variable selection, used in data mining, which is primarily used to minimize the data by eliminating irrelevant and superfluous attributes of any dataset. It also enhances the understandability of data, facilitates better data visualization, reduces training time of learning algorithm or improves the effectiveness of prediction. Prediction in many chronic circumstances such as diabetes, heart disease, strokes, high blood pressure, thalassemia, etc. Various learning algorithms work well and provide more accurate results if more significant and non-resolving elements are stored in the information. Because the medical data set contains several redundant and unnecessary features, and efficient selection method is required to extract disease-specific characteristics. A new selection and classification method was used to develop the approach.

5.2 Feature Selection Approaches
Traditional methods for choosing devices are commonly categorized into 3 groups:
1. Filter Method - This is one of the oldest filtering processes. Filtering functions are performed before implementing a learning algorithm in variable selection using the filter method. This classifies attributes according to certain requirements for assessment. Since it doesn't rely on the classifier, it tends to perform in a variety of ways. The results of these methods are quick and efficient.
2. Wrapper Method - Wrapper methods pick the characteristics by taking into a proper consideration learning algorithm. The main advantage over filters is the most useful features and the best way to pick characteristics of the study algorithm[19]. It also takes into account functional dependency and provides more accurate results compared to filtering approaches.
3. Embedded Method - In an embedded feature selection method, the search usually depends on the learning process. This approach is known as the nesting subset method [22] tests usability of functional subsets & carries out functional selection for training. Normally they use specific study algorithms to maximize the efficiency of a test algorithm.
4. Hybrid Method - Recently, it is one of the widely used approaches for the application of the selection technology employed by researchers. The process incorporates one or more methods to use the benefits of various approaches to produce optimum results. In contrast to wrapping and high device performance in comparison to filter processes, these approaches typically achieve greater accuracy [11].

6 Dimensionality Reduction Techniques
Dimensionality reduction is important for the document because it leads to identifying that set of features(s) which alone shows most variability. Researchers can use this information for using that feature(s) sets for applying various analytical algorithms and thus would reduce the computational processing, memory and time. This may prove to be useful if we involve it for big data processing. There are many studies that have shown how health set data can be grouped in less amount of processing power by reducing the dimensions and working upon the available dimensions [8]. It is also an efficient way to downsize data that can allow efficient storage or recovery of data. The effective use of text mining, pattern matching, image processing, micro-data analysis & image recovery is amongst its applications. Reduction of dimensionality is possible with several algorithms like SVD, SVM, ICA, PCA, etc A descriptive analysis of eight algorithms used for this process has been provided and some are being compared with one another to judge its strengths or weaknesses with respect to its equivalent. Data preprocessing also acts as one of the pivotal procedures in various machine learning projects.

6.1 Algorithms
6.1.1 Singular Value Decomposition (SVD)
SVD is the method of gene selection to decrease data dimensionality. SVD is a factorization approach for the matrix which falls into linear vector algebra. The primary purpose of data analysis on gene expression is to detect and extract the structural constitution inside the data and also for significant gene expression associations.

6.1.2 Principal Components Analysis (PCA)
Dimensionality reduction is a method in data mining that reduces the number of random variables. Dimensionality reduction is commonly defined as a collection of features by choosing a subgroup of all features and an extraction function by merging already existing features and creating a new subset of the combinations. PCA is one of the popular feature extraction techniques.

6.1.3 Support Vector Machine (SVM)
SVM is a newly developed classification technique introduced by Vapnik that has been useful to several domains consecutively. Microarray data of several gene expressions are used for SVM. After many checks, SVM is implemented to identify the tissues of cancer as part of a built-in algorithm. SVM's main aim is to achieve a maximum margin (w) and a real value for data management for an optimal separating hyperplane. The results cannot be linear in reality.

6.1.4 Independent Component Analysis
ICA is a computer technique that is used to separate a different signal into the reduced subcomponents. A common ICA solution is the 'cocktail party problem' where basic voice signals are isolated from test data from people who speak in a room. This complexity is usually interpreted by the lack of delays or echoes. Another important consideration is to take into account the need for N estimates (e.g. microphones) for the mining of primal signals if N sources are present.

6.1.5 Canonical Correlation Analysis (CCA)
In statistics, a way of creating meaning by cross-covariance matrices canonical correlation analysis proposed by Harold Hoteling. Taking into account the two classes of variables and their interrelationships among the variables, canonical analysis of the interrelationships would allow us to discover linear groupings of variables with the highest interrelation. CCA can also be utilized in constructing a model equation that encompasses two sets of variables, e.g. a set of performance measures as well as a set of descriptive variables, or a set of inputs and outcomes.
6.1.6 Locally Linear Embedding (LLE)
Non-linear methods of reducing dimensionalities are generally divided into two categories, for example, those which give a mapping of high space to a lower dimension or vice versa, and the others that only provide a representation of lower-dimensional charts or graphs. In the machine education context, mapping approaches can be viewed as an initial extraction step, followed by pattern recognition algorithms.

6.1.7 Linear Discriminant Analysis (LDA)
LDA is a tool used to resolve the reduction in dimensionality. It is primarily used in the issue of small sample size. This problem can be found in a wide range of data. The amount of current cases in controlled experiments increases comparatively in medical data sets where large amounts of measurements or features are less, or where there are typically significantly greater features or variables than the sample size [21].

6.1.8 Partial Least Squares Regression (PLS Regression)
PLS regression is a mathematical approach that allows some relationship to the main regression of the components, but rather defines hyper-plane variation between the response and independent variables. PLS detects a linear regression model, using the estimated variables and perceptive variables for a different space. [14].

7 LITERATURE REVIEW
J. Cui et al. [2019]It provides a fault diagnosis approach based on the analysis of the main kernel portion and the neural network wavelet. The key nuclear system analysis procedure is used to evaluate the original data parameters of the aero-engine, extract the main characteristics of its component, minimize the parameter size, and create the sample of health status and failure condition with the principal data variable extracted. This is divided into preparing samples and analyzing samples. The neural network wavelet error diagnosis model is based on the training set. Use the test data collection to define and analyze the concept of a diagnostic neural network error. The neural BP network also allows the same set of functional data to be diagnosed. The diagnostic failure pattern for neural network failure is also used for evaluating the initial fault diagnostic technique [15].

K. Wahab et al.[2019]This thesis aims at the development of a key biomedical ontology for the use of UMLS as a data source for chronic liver disease. Because biomedical data and literature are unstructured, ontology can not work on accessible information itself, thus helping to build the structure for existing information. RDF is used as a protégé scheme tool and is structured to explain the principles and their relationship within the chronic liver disease on the basis of these schemes ontology. Field experts evaluate it further [16].

F. Kayaalp et al.[2018]In this research, the patient's laboratory findings and history were used to examine the chronic renal disease. The data set in the study are the data set for chronic kidney disease in the UCI data store. The Bayesian classification algorithm in the chronic renal data set was contrasted with classification algorithms of the neighboring K-\(n\)n (K-nn), obtained from laboratory results of 400 patients, and classification algorithms of Support Vector Machine. Data preprocessing steps like data completion and data standardization have been performed on this data set. In addition, the feature selection was implemented to determine which of the attributes are more important and which will affect the efficiency of classification results before moving on to the classification algorithms for estimation work. The role selection procedure is done through the collection of attribute relief and ratio algorithms [17].

J. G. Teriús-Padrón et al.[2018]Build a modular system to ensure the self-management of chronic diseases in a smart home. Modularity is defined on the basis of a single disease with specific needs for each type of disease: a variety of disease requirements as well as a combination of technologies supporting it in smart homes. In this article, we concentrate on the use of a chronic obstructive pulmonary disease (COPD) patient in an intelligent setting. The proposed solution would allow physical exercise tracking by sensors and photographs, and provide timely feedback on the application, results and environmental conditions which may impact users’ safety [18].

S. Naganjaneyulu and B. S. Rao [2018] a novel feature selection based classification model is proposed to improve the disease classification rate and testing the new type of disease patterns for real-time patient disease prediction. In the proposed model, a novel probabilistic based feature selection measure for classification algorithm is designed and implemented for real-time patient disease prediction using the training datasets. Experimental results show the proposed classification algorithm based on feature selection that is in true positive levels better than conventional algorithms error rate and F-measure are concerned [19].

Y. Cheng et al.[2017] Propose a new method for early detection of chronic diseases using sequential risk patterns mining and classification analysis techniques with time interval knowledge from diagnostic clinical records. The proposed system consists of four phases with a full workflow: preprocessing data, the exploitation of risk patterns, classification modeling, and post-analysis. The methodology can also generate new designs with valuable insights into additional medical studies such as the development and promotion of new markers and therapies, in addition to rich sequencing risk patterns, through a large national clinical database experimental evaluation. This is the first piece of work in our capacity on the issue of time-interval risk dynamics and classification models in the early analysis of chronic diseases. [20].

K. Pavia and B. Srinivasan [2017] Analyze the use in the identification and classification of diseases of filter (F-score) and recursive (Recursive Feature Elimination) algorithms for selection functions. The analysis will also be done with dimensionality reduction algorithms for the Principle Component Analysis. The performance evaluation consisted of three measures of accuracy, flexibility and specificity. Four classifiers, Multilayer Perceptron, Baker Propagation Networks, Help Vector Machines and Extreme Learn machines were used for evaluating the choices of algorithms. The results showed that the F-Score, as well as recidivism disposal, improve diagnosis of thyroid disease by the wrapper-based algorithm were maximum performance, maximum accuracy 98, 14% and ELM classifier. [21].

A. Raji et al. [2016] This paper use wearable sensors to develop real-time surveillance of the vital signs of a patient. Students know vital signs without nursing help from their sensors, and the sensor value is stored as texts by the device. Data mining approaches Train the device for vital sign data. Patients provide the program with their text document, which in effect they know their health without nursing assistance [22].

L. Xiaosong et al. [2015]Proposes the management system for
wellbeing and chronic diseases. The network structure consists of four layers, such as data storage, data service layer, and so on. Six elements of this chronic disease research program include the processing of personal data for patients, the management of personal data in community hospitals, the management, and management of personal data in hospitals, the management of patient physiological information, the management of patient status review reports and the management of function and system protection. Increasing population aging in our country means a significant increase in the number of chronic disease patients. The majority of chronic diseases cannot be properly treated and will affect the family and society of the patient. [23].

8 CONCLUSION
Prediction of chronic diseases plays a key role in computer health care. It is important that the disease is diagnosed early. This paper provides an analysis of various methods of selection or reduction of dimensionality used to predict and classify these conditions. We have discussed the importance of feature selection methodologies for improving the accuracy and performance of classification systems. Also, define the concept of principal component analysis statistical method. A dimension reduction technique can help to solve the issue of “dimensionality curse” as large data sets mean the huge number of records and an even greater number of features and also enhance its computational efficiency.

REFERENCES

DOI:10.1016/j.eij.2018.03.002
[23] L. Xiaoong, S. Gang and C. Yong, “Design and
Development of the Web-Based Health and Chronic Disease Assessment Management System,” 2015 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, 2015, pp. 72-75. doi: 10.1109/ITME.2015.167.