International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 7 Issue: 1 17 - 20

# Various Feature Selection Techniques in Type 2 Diabetic Patients for the Prediction of Cardiovascular Disease

Dr. P. Radha
*radhasakthivel09@gmail.com*

**Abstract**: Cardiovascular disease (CVD) is a serious but preventable complication of type 2 diabetes mellitus (T2DM) that results in substantial disease burden, increased health services use, and higher risk of premature mortality [10]. People with diabetes are also at a greatly increased risk of cardiovascular which results in sudden death, which increases year by year. Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. Usually medical databases of type 2 diabetic patients are high dimensional in nature. If a training dataset contains irrelevant and redundant features (i.e., attributes), classification analysis may produce less accurate results. In order for data mining algorithms to perform efficiently and effectively on high-dimensional data, it is imperative to remove irrelevant and redundant features. Feature selection is one of the important and frequently used data preprocessing techniques for data mining applications in medicine. Many of the research area in data mining has improved the predictive accuracy of the classifiers by applying the various techniques of feature selection This paper illustrates, the application of feature selection technique in medical databases, will enable to find small number of informative features leading to potential improvement in medical diagnosis. It is proposed to find an optimal feature subset of the PIMA Indian Diabetes Dataset using Artificial Bee Colony technique with Differential Evolution, Symmetrical Uncertainty Attribute set Evaluator and Fast Correlation-Based Filter (FCBF). Then Mutual information based feature selection is done by introducing normalized mutual information feature selection (NMIFS). And valid classes of input features are selected by applying Hybrid Fuzzy C Means algorithm (HFCM).

***Keywords***- *Data mining, Feature selection, data preprocessing, Symmetrical Uncertainty Attribute set evaluator, FCBF, NMIFS, HFCM.*

*****

## I. INTRODUCTION

The diagnosis of heart disease mainly depends on complex grouping of clinical and Pathological data. Due to this complexity, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is very crucial. Numeral number of tests must be requisite from the patient for detecting a disease. Machine learning based method is used to classify between healthy people and people with disease. Cardiovascular disease is the principal source of deaths widespread and the prediction of Heart Disease is significant at an untimely phase. In order to reduce number of deaths from heart diseases there has to be a quick and efficient detection technique.

Feature subset selection is a preprocessing step in the machine learning. Feature selection plays an important role in identifying and removing more irrelevant and redundant information which increases learning accuracy. [8]. Feature selection is often considered as a necessary preprocess step to analyze the data, as this method can reduce the dimensionality of the datasets and often conducts to better analysis [9]. Research [10] shows that the reasons for feature selection include improvement in performance prediction, reduction in computational requirements, reduction in data storage requirements, reduction in the cost of future measurements and improvement in data or model understanding.. Feature selection techniques identify the features that mostly improve the predictive accuracy of the classifiers.

Learning accuracy increases by reducing the dimensionality and removing the features that are useful in predicting class. Features can be irrelevant data. It refers to the problem of identifying the important features discrete, continuous or nominal. Generally features are of three types. They are Relevant, Irrelevant and Redundant. Feature selection methods wrapper and embedded models. Filter model rely on analyzing the general characteristics of data and evaluating features and will not involve any learning algorithm, where as wrapper model uses the determined learning algorithm and uses these learning algorithms to perform the provided features in the evaluation step to identify relevant feature. Embedded models incorporate feature selection as a part of the model training process.

Data from medical sources are highly voluminous nature. Many important factors affect the success of data mining on medical data. The training phase becomes more difficult to knowledge discovery if the data is irrelevant, redundant.

### Related work

Few research works has been carried out for diagnosis of various diseases using data mining. This approach is to apply feature subset selection to enhance the prediction of heart disease in type 2 diabetic patients. M.A. Jabbar et.al proposed a new algorithm combining associative

classification and feature subset selection for heart disease prediction [5]. They applied symmetrical uncertainty of attributes and genetic algorithm to remove redundant attributes. Enhanced prediction of heart disease using genetic algorithm and feature subset selection was proposed by Anbarasi et.al [10]. Heart disease prediction using associative classification was proposed by M.A. Jabbar et.al [11]. They combined maximum clique concept in graph with weighted association rule mining for disease prediction. Feature subset selection using FCBF in type II Diabetes Databases was proposed by Sarojini Balakrishnan et.al. [17]. Hybrid Artificial Bee Colony Algorithm with Differential Evolution Based Feature Selection and Semi supervised Learning Prediction Model For the Risk Of Cardiovascular Disease in Type-2 Diabetic Patients was proposed by P. Radha et.al [18].

**Proposed System**

Feature selection has been widely applied in a number of branches of computer science including computer vision, pattern recognition, classification and machine learning. The technique of feature selection in medical databases enables to find small number of informative features leading to potential improvement in medical diagnosis.

**Dataset Information**

The dataset collected from real patient records which includes the following attributes for diabetes patients records Plasma glucose concentration a 2 hours in an oral glucose tolerance test , Diastolic blood pressure (mm Hg) ,Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml) ,Body mass index (weight in kg/(height in m)^2) ,Diabetes pedigree function ,Age (years) ,Class variable (0 or 1).These data are collected with the following CVD risk factors which includes BMI (Body Mass Index) , Weight (kg) ,Waist circumference (cm) , Systolic blood pressure (SBP) (mmHg) , Diastolic blood pressure (DBP) (mmHg) ,Glucose (mg/dl) ,Total cholesterol (mg/dl) , High-Density Lipoprotein cholesterol (HDL-c) (mg/dl) , Low-Density Lipoprotein cholesterol (LDL-c) (mg/dl) ,Triglycerides (mg/dl) ,HbA1c (glycosylated hemoglobin) (%) Fibrinogen (mg/dl), ultrasensitive C reactive protein (us-CRP) (mg/L). If the value of each and every attributes values are changed to analysis the risk factor of CVD for type 2 diabetes (T2D). Managing the numerous risk factors responsible for CVD in T2D represents an ongoing challenge for primary care clinicians, strongly influencing their decisions about treatment approaches for this complex disease.

Artificial Bee Colony (ABC) is one of the most recently used algorithm is motivated by the intelligent behavior of honey bees. It is as like Particle Swarm Optimization (PSO) and Differential Evolution (DE) algorithms and uses only common control parameters such as colony size and maximum cycle number. The optimization tool is ABC that provides a population based search procedure in which individuals called foods positions are modified by the artificial bees with time and the bee's aim is to discover the place of food sources with high nectar amount and finally the one with the highest nectar.

One of the main swarm based approaches is Artificial Bee Colony (ABC) algorithm. This algorithm recommends the intellectual searching behavior of a honey bee swarm. The dependency of artificial bees includes three major clusters of bees. They are employed bees, onlookers and scouts. To surpass the initial random population values for features from T2D patient records with CVD risk factors, Differential Evolution (DE) methods are used to generate initial population values for each feature in T2D patient records data with CVD risk factors. Features are chosen depending on the dancing area to select the best feature is known as onlooker bee and the one going to the feature selection of T2D patients feature source visited by it before is named as employed bee. The scout bee performs arbitrary investigation for determining new feature selection source. During the initial phase of Artificial Bee colony the feature selection randomly generates initial population of size SN, where SN is the overall number of input T2D patients feature samples with attribute measurement results from kernel density estimation for CVD risk factor investigation indicates the size of the population. Every feature selection solution represents a D-dimensional vector. D is the number of features in the T2D patients with CVD risk factors.

The differential evolution algorithm takes mutation operation as initial operation to create new feature population for T2D samples and selection task to manage the feature selection task toward the CVD risk factor assessment. The first feature selection method used is Differential evolution and Artificial Bee Colony to select the significant feature from T2D patient to predict the risk assessment of cardiovascular diseases. The differential evolution algorithm does the mutation operation to create new feature population for T2D patient samples. This algorithm also exploits a non-uniform crossover that can take child feature vector parameters from one feature vector T2D patient's records with CVD risk factors typically than it does from others.

After exploiting the existing features of T2D patient population, the recombination i.e. crossover operator competently mixes up T2D patient features information regarding successful combinations. Every generated feature samples is accomplished depending on maximum number of cycles, $C= 1,2,3....MCN$ for bees.

The next feature selection is a filer based feature selection approach symmetrical Uncertainty Attribute set selector and Fast Correlation Based Filer (FCBF). These two methods are used to remove both irrelevant and redundant features.

Given N samples of data set with n features and a class C, the feature selection problem is to find from the M-dimensional observation space, $S^M$, a subspace of m features, $S^m$. The total number of subspaces is $2^M$.

The feature selection process consists of two phases. During the first phase Symmetric Uncertainty (SU), the measure of correlation between the feature i and the class C is calculated for each feature. During the second phase $S_{best}$ is further processed to remove the redundant features. Starting from the first feature the redundant features are eliminated and only the predominant one are kept among all the selected relevant features.

**Feature Selection using Mutual Information Feature Selection**

A filter/wrapper hybrid feature selection method is proposed that has two parts: a genetic algorithm and a neural network classifier. A GA called deterministic crowding (DC) is used. In DC, all individuals in the population are randomly paired and recombined, i.e., probability of crossover is one. The binomial crossover is used here because it has no positional bias. Mutation is optional in DC. The resulting offspring has a tournament with its nearest parent in terms of Hamming distance. The winners are copied to the new population for the next generation. For the feature selection problem, subset of selected features are represented as bit strings of length L (total number of features in the problem at hand), where "1" in the th position indicates that the ith feature is included in the subset, and "0" indicates that the ith feature is excluded. In order to evaluate the fitness of an individual (chromosome), the corresponding binary string is fed into an MLP classifier. The size of the input layer is fixed to L but the inputs corresponding to non selected features are set to 0. The fitness function includes a classifier accuracy term and a penalty term for a large number of features. The accuracy of the classifier is measured as the maximum rate per unit of correct classifications on a validation set. BPQ is faster than first-order algorithms and many second-order algorithms such as Broydon–Fletcher–Goldfarb–Shanno (BFGS). (15)

A method to initialize the initial GA population with good starting points that makes use of the feature ranking delivered by NMIFS is introduced. In addition, a new mutation operator guided by NMIFS is used to speed up the convergence of the GA. This operator allows adding a relevant feature or eliminating an irrelevant or redundant feature from individuals in the GA population. The mutated individual is evaluated first to verify whether mutation improves the classifier accuracy. Only if the mutant's fitness obtained is better than that of the original individual the mutation is completed, otherwise the mutated feature is restored. This is the only mutation operator used here. The proposed mutation operator can be used with any classifier, not necessarily an MLP neural network, since the mechanism for accelerating the convergence does not depend on the nature of the classifier.

**Unsupervised Learning using Hybrid Fuzzy C Means Clustering**

As the first step, before the application of the Classification algorithms, validating the chosen classes using the unsupervised methods is aimed. This work uses an HFCM clustering to validate the preprocessed dataset, and then assign class labels to similar cluster, the clustering algorithm. The problem associated with fuzzy c-means is the number of clusters to be generated for the given dataset needs to be specified, this can be solved by this proposed method. In this method, the fuzzy c-means combined with the Quantum PSO algorithm provides the statistical frame work to model the cluster structure of gene expression data. It makes use of probabilistic models which can explain the probabilistic characteristics of the given systems and helps to find the precise number of clusters for the given dataset so that the resultant value of QPSO can be used as number of clusters k. The main objective of using this hybrid method is to minimize the objective function value in fuzzy c-means. In this approach, each particle is a D dimensional candidate solution in one of the C clusters that can be formally represented as the matrix X: (9). A population of particles is randomly initialized and personal as well as global best positions are determined. Subsequently membership values are computed and a cost is assigned to each particle. The QPSO algorithm minimizes the cost associated with the particles through recursively calculating the mean best position using eq. (7), the membership values and cost function through eqs. (1) and (3) and updating the candidate cluster centre solution X. The algorithm is terminated if there is no improvement in the global best and the algorithm stagnates or if the preset number of iterations is exhausted. By using the stochastic and non-differentiable objective function handling capabilities of QPSO within the FCM algorithmic framework, the problem of stagnation in local minima within a multidimensional search space is mitigated to an extent better than that possible with only the traditional FCM.(15) The pseudo code of FCM QPSO is outlined below:

**Algorithm**

1: for each particle xi
2: initialize position
3: end for
4: Evaluate membership values using eq. (1)
5: Evaluate cost using eq. (3) and set pbest, gbest
6: do
7: Compute mean best (mbest) position using eq. (7)
8: for each particle xi

9: for each dimension j
10: Calculate local attractor Φij using eq. (8)
11: if k ≥ 0.5
12: Update xij using eq. (9) with '+'
13: else Update xij using eq. (9) with '-'
14: end if
15: end for
16: Evaluate cost using eq. (3) and set pbest, gbest
17: end for
18: while max iter or convergence criterion not met

## II. CONCLUSION

Feature subset selection is a preprocessing step used to reduce dimensionality and to remove irrelevant data. It also increases accuracy which increases accuracy thus improves the predictability of the problem. The reasons for feature selection includes improvement in performance prediction, reduction in computational requirements, reduction if data storage requirements, reduction if the cost of future measurements and improvement in data or understanding the model. This also improves the predictive accuracy if the classifiers by applying the techniques of feature selection. In this proposed work Artificial Bee Colony with Differential evolution (ABC with DE), symmetrical Uncertainty Attribute set selector and Fast Correlation Based Filer (FCBF) are used to select the fittest feature from type 2 diabetic patients for the prediction of cardiovascular disease. Then Mutual information based feature selection is done by introducing normalized mutual information feature selection (NMIFS) and valid classes of input features are selected by applying Hybrid Fuzzy C means algorithm (HFCM). The experimental results show that the accuracy of the classifier has improved in prediction after applying feature subset selection.

## References

[1]. Jackson J.E., 1991, A User's Guide to Principal Components, New York, John Wiley and Sons.
[2]. Jin W Z and Patti, E M., 2009, Genetic determinants and molecular pathways in the pathogenesis of type2 diabetes, Clinical Science, 116: 99-111.
[3]. Jolliffe I T., 986, Principal Component Analysis, Second edition, Springer - Verlag, New York.
[4]. Kullback, S., 1987, Letter to the Editor: The Kullback–Leibler distance, The American Statistician, 41 (4): 340–341.
[5]. Kuller L H., 1995, National Diabetes Data Group. Stroke and diabetes. In: Diabetes in America. Bethesda, Md: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 449–456.
[6]. Larose D.T. and Wiley J., 2006, Data mining methods and models, Wiley Online Library. Lazar, M A., 2005, How obesity causes diabetes: not tall tale? Science, 116: 99-111.
[7]. Malecki., Maciej T., 2005, Genetics of type 2 diabetes mellitus, Diabetes Research and Clinical Practice, 68S1: S10- S21.
[8]. Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998
[9]. P. Radha, Dr. B. Srinivasan, "Feature Selection Using Particle Swarm Optimization for Predicting the Risk of Cardiovascular Disease in Type-II Diabetic Patients", COMPUSOFT, An International journal of advanced computer technology, Volume 3, Issue 11, November-2014.
[10]. Sarojini Balakrishnan, Ramaraj Narayanaswamy, Feature Selection Using FCBF in Type II Diabetes Databases, International Conference on IT to Celebrate S. Charmonman's 72nd Birthday, March 2009, Thailand.
[11]. M. Ambarasi et.al, "Enhanced Prediction of heart disease with feature subset selection using genetic algorithm", JEST Vol2 (10) pp 5370-5376(2010).
[12]. M A. Jabbar et.al," Knowledge discovery using associative classification for heart disease prediction", AISC 182 pp29-39, Springer-Verlag (2012).
[13]. MA.Jabbar et.al, "Knowledge discovery from mining association rules for heart disease prediction", JAJIT, Vol 41(2) pp 45-51 (2012).
[14]. Sellappan et al., "Intelligent heart disease prediction system using data mining techniques", IEEE (2008).
[15]. Radha P, " Normalization based Intelligent risk factor classification system for type-2 Diabetic patients", Journal of advanced research in Dynamical & Control Systems, Vol 10, 05- special issue, 2018