Hybrid PCA-GA Recursive Method for gene analysis of TCF7L2 Associated with T2DM

K.Vijayalakshmi[1]
[1]Asst.Professor,Dept.of Computer Science, Sri Venkateswara University, Tirupati.

Abstract: Various methodologies have been developed for investigating the multiple complex diseases. DNA microarray has attracted huge attention of researchers due to its significant nature to provide the efficient information about patient. DNA microarray is widely used for investigation of obesity, neurodegenerative disease, cancers and diabetes etc. Microarray technology has grown rapidly in the bio-medical field and applications. In order to understand the gene expression data for T2DM, normal and abnormal gene expression data are collected. In the data base most of the genes are considered in this case are not directly associated with disease and can be denoted as redundant which increases the complexity and requires more time for computation. To overcome this, Dimensionality Reduction is implemented on the data set.

Keywords: T2DM, Dimensionality Reduction, Feature Reduction, TCF7L2 gene.

I. INTRODUCTION

In Diabetes disease, gene identification is a crucial phase. Additionally, the vast majority of the genes taken into account for the review are not vital and thus, they are redundant. Microarray database contains huge number of genes as data which are very high when compared with low sample size data resulting in increase in gene analysis complexity. Dimensionality reduction is the process of reducing the number of random variables in consideration, via obtaining a set of principal variables. Such expands the computational overload and consumes large amount of time, particularly in situations where the aim is to recognize the most oppressive gene which may be the reason behind the illness. In spite of the fact that these gene selection strategies have the ability to eradicate the unusable genes, they are not very useful in removing repetition of genes in microarray. This elevates the cost of calculation and lowers the precision of data recognition and classification. The identification of the true function of gene and other related useful data is very troublesome because of the low sample data in microarray, and noise present in them. The problem of redundancy can be resolved by consolidating the gene expression information.

II. DISCUSSION OF EXISTING APPROACH

Support Vector Machine (SVM) is a machine learning algorithm implemented for time series forecast and classification problems. It is generally connected in the biological science domain, particularly in Bioinformatics. SVM [41] has the ability to deal with nonlinear classification problems productively by mapping the data examples onto a higher dimensional space, with the help of a nonlinear kernel function. The SVM is model-free approach, which driven by data. This property of SVM makes it an essential segregating power in classification process. In the existing work the author modified SVM as Support Vector Machine Recursive Feature Elimination (SVMRFE) [14] is for reducing dimensionality. SVMRFE was utilized to recognize and distinguish highly biased target gene in four distinctive microarray example data sets for type II diabetes. These examples have been obtained from the DGAP (Diabetes Genome Anatomy Project) and GEO (Gene Expression Omnibus database). The following steps are used for existing SVMRFE:

Input : Training samples $X_0 = [x_1, x_2, x_3 \dots x_n]^T$ Class labels (1 for normal or 0 for diseased) $y = [y_1, y_2 \dots y_n]^T$
Initialize: Surviving genes $s = [1,2,3 \dots n]$ Gene ranking list $r = []$ Limit training samples to good genes $X = X_0(:, s)$ Apply classifier training process $\alpha = SVN - train(X, y)$ Compute the weight from each selected gene: $w = \sum_k \alpha_k y_k x_k$ where k indicates the k^{th} training pattern Compute the ranking criterion for the i^{th} gene $R(i) = (w_i)^2$ Mark the gene with the lowest ranking $g = arg\,min\,(R)$

Renew the gene-ranking list
$r = [s(g), r]$
Eliminate the gene with the lowest ranking
$s = s(1: g - 1, g + 1: length(s))$
Repeat until s = []
Output:
A gene-ranking list r

III. DATASET DESCRIPTION

For the proposed work author gathered the information from NCBI repository with population size of 5900 where 23 variables are associated with this dataset.

Dataset Name: Type2Diabetes_NHS_HPFS_Phenotype
Dataset Accession: pht000115.v2.p1
Dataset Accession: pht000115.v2.p3
Table Description:

S.No	Variable Name	Variable description
1	GeneId	GENEVA identification number
2	BMI	BMI in kg/m2
3	hisp	Hispanic ethnicity
4	Case	Diabetes case status
5	race	Race variable for NHS
6	magn	Magnesium intake
7	Age	Age in years
8	trans	Transfat intake
9	Race2	Race variable for HPFS
10	SMK	Smoking habit
11	pmh	Menopausal status and hormone use
12	Act	Total physical activity
13	Ht	Height in meters
14	heme	Heme iron intake
15	Alcohol	Alcohol intake
16	ceraf	Cereal fiber intake
17	Pofa	Poly-saturated fat
18	Woman	Sex
19	Gl	Glycemic load
20	Wt	Weight in kg at time of blood draw
21	famdb	Family history of diabetes among first degree relatives
22	Hbp	Reported high blood pressure during or before blood draw
23	Chol	Reported high blood cholesterol during or before blood draw

Table 1: Type2Diabetes_NHS_HPFS_Phenotype Data base

IV. DESIGN AND DEVELOPMENT OF HYBRID PCAGA- RECURSIVE FEATURE REDUCTION METHOD

The author introduced a consolidated strategy for dimensionality reduction where selection of feature reduction with classification approach is implemented to generate forecasted outcomes. Principle Component Analysis (PCA) method have been used for gene selection in order to eradicate the less useful and repetitive genes. The dimensionality is lowered by means of forecasting the high-dimensional information into single dimensional information. It is then followed by classification of information in the one-dimensional space. According to PCA, the normal estimation of genes that demonstrates a major contrast between the two classes with respect to their variance, tend to get a larger value.

For this author proposed a new Hybrid PCAGA- Recursive Feature Reduction (PCAGA-RFR) Method by combining principle component classification method, genetic algorithm to achieve optimized feature reduction. These feature set are passed on to data mining algorithm, which results in classification of the data. Moreover, group prediction is implemented only for the multiple groups that are characterized with numerous genes. The procedural steps are as follows

Step 1: Initialize Input set with attributes
X={Idg,BMI,hispace,magn,age,trans,race2,smk,pmh,act,ht,heme,alcoh…hbp,chol}
Assumed as $X_n = \{X_{n1}, X_{n2} \ldots, X_{nF}\}^T$ in the computational algorithm where $= 1\ to\ 5900$; F is number of attributes=23
Step 2: Repeat PCA for feature reduction until to get reduced dimension set for any given input data set
Step 2.1: Computation of covariance matrix i.e. $F \times S$ matrix by adjusting feature vector I_n.
Step2.2: Decompose the covariance matrix C into is a diagonal matrix D which can be expressed as $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_F)$.
Step2.3: Eigen vector formed matrixΛ, the original data vector I_n is transformed to an uncorrelated vector y_n and sort Eigen
values in descent order and discard the smallest or less representative ones, this leads to truncation of the matrix K.
Step 3: After getting reduced dimension set apply genetic algorithm on reduced data set K for feature selection and optimization.
Step 4: The optimized feature set obtained from step 3 which is used for classification analysis.

In Step 3, genetic algorithm implemented for feature selection and optimization and explained in the following sections

4.1 Implementation Genetic Algorithm for feature selection and optimization:
A large number of features are mostly generated by using terms extracted from the given corpus. In Genetic Algorithms, it is too difficult to perform feature selection due to its runtime complexity and iteration. Therefore, we haven't done feature selection in the single GA process. That is, we have generated the partial features of fixed length through the GA process and grouped each partial feature to obtain a final feature set.
The feature with a high frequency can be very important so that the repetitive process is allowed. The length of the final feature set is determined in proportion to the fixed length of features generated from the corpus by the following iterative steps

Proposed Feature Selection Algorithm
Feat_length= Number of total features in database Sub_Feat= Total number of sub-feature available Final-Feat= Final selected feature set. Max_Gen= Maximum value of generation. Time_limit= Maximum limit for elapsed time Pop_Size= Size of Population

// Chromosome Length = sub features
// Gene vale = feature index(1<indx<Feat_length)
Initialize
While (Final_Feat> Final_Feat_Length) {
P<-Initial_Population (Feat_Length, Sub_Feat, Pop_Size)
Fitness_Compute(p)
Initial_Generation =1
While (max-gen>= generation OR lim_time>Elapsed-Time)
{
New_Pop<- Rank_Selection (p)
Apply_Crossover(New_Pop)
Apply_Mutation(New_Pop)
p (population) <- New_Pop
Fitness_Compute(p)
Elapsed Time =Elapsed Time+ Current Time
Generation = generation +1
}
Sub_Feat <- Top_Fitness_Feat (p)
Final_Feat <-Add_Feat(Sub-Feat)
}
return Final_Feat

This final feature set is known as optimized feature set which is used for classification analysis of various techniques like Neural Networks, Decision Trees, KNN Classifier, SVM classifiers.

V. RESULTS AND DISCUSSION

This section provides a complete experimental study and performacne analysis for T2DM prediction. As dicussed in previous section that proposed data mining approach uses data pre-processing, feature selection, feature reduction for predicting the diabetes in given microarray database. In order to perform this operation, we have considered gene microarray data where various attributes are given for each user. Along with this to prove TCF7L2 is highly related to T2DM we segregated data into two groups where first group contains TCF7L2 gene, second group does not contain TCF7L2 gene.This training and testing process is divided into two parts: On the above categorized dataset, we applied the proposed Hybrid PCAGA- RFR Method and observed the various measurement metrics such true positive rate, false positive rate, accuracy, confusion matrix, precision and recall curve analysis, ROC curve analysis.

5.1 T2DM Associated gene analysis:

In this case, we have considered the reduced data set for training and testing which obtained by implimenting above proposed method. The following attributes are present in the reduced feature subset {1.Magnesium intakes, 2. Family history of diabetes among first degree relatives, 3.Hispanic ethnicity 4. Menopausal status and hormone use, 5. Age in years}. Then we implimented various classification schemes for performance evlauation such as Decision Tree classfier, KNN classification, Neural Netwrk and Support Vector Machine.

5.2 Classification Accuracy Performance:

Given input database contains two class labels such as {true, flase}. This data label is used for classification. Here we present classification accuracy performance for Case-I where TCF72 gene is present.

	Decision Tree	KNN Performance	SVM Performance	Neural Network Performance
Classification Accuracy	84.54	85.14	97.99	97.59

Table 2: Classification Accuracy performance for Case-I

Above given table gives classification accuracy performances by using various classification schemes. According to this performacne analysis,support vector machine and neural network classification schemes obtians better classification accuracy when compared with decision treen and KNN classification approach. This performance is evlauated using TCF7L2 gene that has impact to the T2DM.

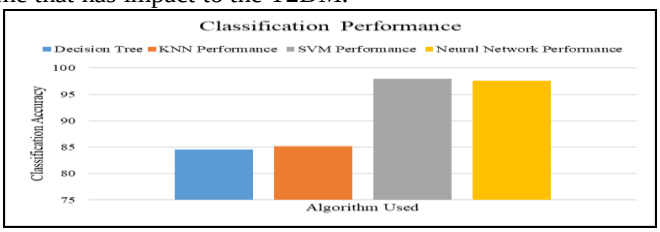

Figure1: Classification Performance

The above figure shows a combined performance analysis for T2DM prediction form the given dataset. This analysis shows that SVM and Neural network obtains better classification perforamcne. Furthermore, we are presenting statistical performances of differnet classifiers interms of confusion matrix, specificity, senstivity, false score, precision and recall.

5.3 Statistical Performance with various parameters:

5.3.1 Confusion Matrix

Confusion matrix is measurement of correctly classified and misclassified intstances under the specified class laebl. Table 3 shows confusion matrix of decision tree classifier.

Class 1 (Non-Diabetic)	11	38
Class 2 (Diabetic)	36	413

Table 3: Confusion matrix by Decision Tree classifier

Similarly, Table 4,5 and 6 shows confusion matrix of KNN, SVM and Neural Network classifiers.

Class 1 (Non-Diabetic)	13	36
Class 2 (Diabetic)	14	435

Table 4: Confusion matrix by KNN classifier

In this anlaysis, KNN classifier is able to predict 413 instances as true positive in class 2 whereas true negative instnaces are predict as 38.

Class 1 (Non- Diabetic)	48	1
Class 2 (Diabetic)	0	449

Table 5:Confusion matrix by SVM classier

Support vector machiine is well-known classification shceme which is also appleid here. Confusin matrix shows a significan performance by avoiding true negative and false positve classified instances.

Class 1 (Non-Diabetic)	49	0
Class 2 (Diabetic)	12	437

Table 6:Confusion matrix by NeuralNetwork classifier

5.3.2 Senstivity Performance

Moreover, we present perforamcne analysis by considering senstivity performance for each class measured with predictd output.

Classification Scheme	Sensitivity (Class-1)	Sensitivity (Class-2)
Decision Tree	0.2653061	0.9688196
KNN	0.224489796	0.919821826
SVM	0.979591837	1
Neural Network	1	0.973273942

Table 7: Senstivity Performance for Classifier

Above given Table7 presents sensitivty performance measurement for T2DM prediction using different classifiers. This analysis shows that all the classifiers are more senitive towards class 2 which is assigned as diabeteic class.

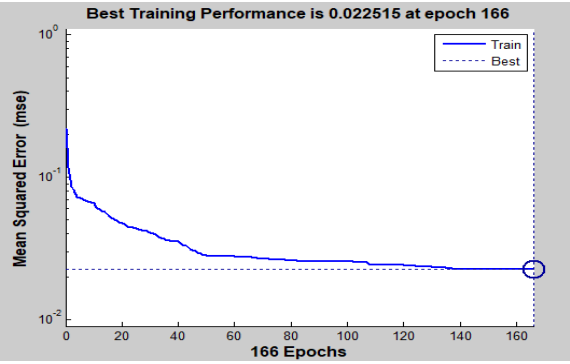

Figure 2: Training Performance through NN

Above given figure shows neural network perforamnce analysis during prediction state. We have conisdered total 1000 number of epochs whereas figure shows that at 166 th epoch, neural network obtains the reduced error.

5.3.3 Specificity, Precison, Recall Perforamnce

Another perforamnce parameter is conisdered as specificity which is depicted in table 8,9,10. Specificity is the measurement of true positive instance predicition for T2DM having TCF72 gebediabetic and non-diabetic class.

Table 8: Specificity Performance

Classification Scheme	Precision (Class-1)	Precision (Class-2)
Decision Tree	0.4814815	0.2653061
KNN	0.224489796	0.915742794
SVM	1	0.997777778
Neural Network	0.803278689	1

Table 9: Preicison Performance

Classification Scheme	Specificity (Class-1)	Specificity (Class-2)
Decision Tree	0.9688196	0.2653061
KNN	0.919821826	0.224489796
SVM	1	0.979591837
Neural Network	0.973273942	1

Table 10: Recall Performance

Classification Scheme	Precision (Class-1)	Precision (Class-2)
Decision Tree	0.2653061	0.9688196
KNN	0.234042553	0.919821826
SVM	0.979591837	1
Neural Network	1	0.973273942

Furthermore, we present graphical representation of each classification scheme by ploting ROC curve, Precision-Recall
(P-R)curves. Given below Figure3,4 shows graphical representation of performacne anlaysis using various classification schemes by measuring the ratio of true positive rate and false positive rate.

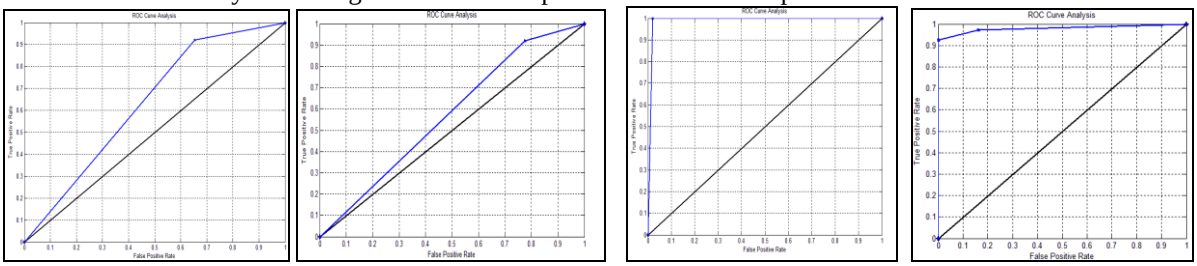

Figure 3: ROC curve analysis by using Different Classifiers

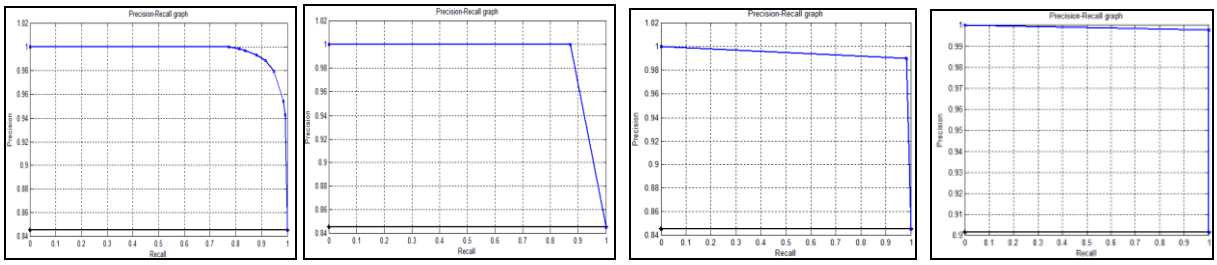

Figure 4:P-R curve analysis by using Different Classifier

In order to validate the assoaciation of TCF7L2 gene with T2DM we applied different classification schemes on dataset by implenting Hybrid PCA-GA RFR dimensionality redction technique.This study shows that SVM and NN are more specific to the TDM having TCF7L2 gene.

VI. CONCLUSION

In this chapter, author proposed new Hybrid PCAGA- RFR to lower the dimensionality of the dataset. On the obtained reduced attribute data set we applied different classification schemes to predict the most biased gene for contributing to T2DM. As a result we find that more number of T2DM predictions taken place in database contain TCF7L2 gene. From this analysis, it can be concluded that TCF7L2 gene (variant rs7903146) has more association and serious impact on T2DM.

VII. REFERENCES:

[1] Cauchi, Stéphane, et al. "TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: global meta-analysis." Journal of molecular medicine 85.7 (2007): 777-782.
[2] Common Variants in the TCF7L2 gene are strongly associated with Type 2 Diabetes Mellitus by G.R. Chandak. C.S.Janipalli et al. in Springer
[3] Choosing SNPs Using Feature Selection by Tu Minh Phuong, Zhen Lin Russ B.Altman in Proceedings of the IEEE Conference
[4] Data mining and Genetic algorithm based gene/SNP selection by Shital C.Shah,Andrew Kusiak in Elsevier publication

[5] Dhawan, Dipali, and Harish Padh. "Genetic variations in TCF7L2 influence therapeutic response to sulfonylurea as in Indian diabetics." Diabetes Research and Clinical Practice 121 (2016): 35-40.
[6] Development of a Predictive Model for Type 2 Diabetes Using Genetic and Clinical Data[60] by J Lee, B Keam et.al, Public Health Res Perspect 2011 2(2), 75e82,pISSN 2210-9099 ISSN 2233-6052
[7] E. AbuKhousa, P. Campbell "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems"2012 International Conference on Innovations information technology IIT 2012
[8] E. J. Parra, L. Cameron, L. Simmonds et al., "Association of TCF7L2 polymorphisms with type 2 diabetes in Mexico City," Clinical Genetics, vol. 71, no. 4, pp. 359–366, 2007.
[9] Genetics of Diabetes by Richard B. Horenstein and Alan R. Shuldiner niversity of Maryland School of Medicine, Division of Endocrinology, Diabetes and Nutrition, 660 West Redwood Street, Room 494, Baltimore, MD 21201, USA.
[10] Gene Expression Analysis of Type-2 Diabetes with Parental History –A computational Approach by V.Chandra Sekar, Prof.P.Srinivas Rao Pathomechanisms of Type 2 Diabetes Genes by Harald Staiger,Fausto Machicao in Endocrine Reviews.