International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 6, December 2015, pp. 1516~1524
ISSN: 2088-8708

Feature Reduction in Clinical Data Classification using Augmented Genetic Algorithm

Srividya Sivasankar, Sruthi Nair, M.V. Judy
Department of Computer Science & I.T, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham

Article Info

Article history:

Received Apr 27, 2015
Revised Aug 11, 2015
Accepted Aug 30, 2015

Keyword:

Classification
Feature reduction
Genetic algorithm
PCA

ABSTRACT

In clinical data, we have a large set of diagnostic feature and recorded details of patients for certain diseases. In a clinical environment a doctor reaches a treatment decision based on his theoretical knowledge, information attained from patients, and the clinical reports of the patient. It is very difficult to work with huge data in machine learning; hence to reduce the data, feature reduction is applied. Feature reduction has gained interest in many research areas which deals with machine learning and data mining, because it enhances the classifiers in terms of faster execution, cost-effectiveness, and accuracy. Using feature reduction we intend to find the relevant features of the data set. In this paper, we have analyzed Modified GA (MGA), PCA and the combination of PCA and Modified Genetic algorithm for feature reduction. We have found that correctly classified rate of combination of PCA and Modified Genetic algorithm higher compared to the other feature reduction method.

Corresponding Author:

M.V Judy,
Department of Computer Science & I.T,
Amrita School of Arts & Sciences,
Amrita Vishwa Vidyapeetham.
Email: judy.nair@gmail.com

1. INTRODUCTION

In a clinical environment, a doctor makes a medical diagnosis based on his medical expertise, symptoms of a patient, and from the patient's test reports. Medical diagnosis is a critical task which involves high precision and with no chances of errors. Redundancy in hospital records, negligence of other medical conditions, ambiguous responses from the patient, can result in a delay in diagnosis or perhaps even a wrong diagnosis. To improve the accuracy of diagnosis for effective treatment, we propose a machine learning algorithm to analyze whether a patient tests is positive or negative for a certain disease. Since the data we deal with is too large and complex, we have to first reduce the data using feature reduction techniques like, Genetic Algorithm (GA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA) etc. The most common techniques among these are GA and PCA. Feature Reduction is the process of removing redundant or irrelevant data from the original dataset. With the help of feature reduction the execution time of classification is considerably reduced, and an increased accuracy rate is obtained due to the removal of redundant and noisy data. Retaining all the unwanted attributes during the training process consumes a lot of memory, storage space and CPU resources, to overcome this problem feature reduction is performed. Clinical Data Sets define a standard set of information that is generated from care records, organization or system that captures the data. Clinical data are a primary resource for most of the health and medical research. Clinical data can be collected either during the course of ongoing patient care or it can be through formal clinical trial.

Clinical data fall into the following categories:

1) Electronic health record
2) Administrative data
3) Claims data
4) Disease data
5) Health survey
6) Clinical trial data

In our paper, we deal with an electronic health record which consists of medical details of the patient. Clinical data are a combination of different attribute types. While comparing clinical data with a normal data we observe that, the latter is mostly composed of a single data type while the former is a combination of attribute types.

In recent decades, considerable research has been done in machine learning and data mining techniques. Among these GA is found most useful in medical knowledge discovery. GA is a search methodology which is developed using the principle of natural selection. In these papers, genetic algorithm was used to identify the key diagnostic features and classify whether the patient is suffering from a particular disease or not [1], [2]. Jin-Xing Hao, Yan Yu, Rob Law and Davis Ka Chio Fong proposed genetic algorithm based learning approach to understand customer satisfaction [5]. Abdulhamit Subasi, M. Ismail Gursoy proposed a PCA feature reduction system and used this data for EEG classification [6]. In the proposed system, Genetic Algorithm chromosome encoding is done in value format. We also use PCA and GA for feature reduction, PCA is used for the identifying the pattern in the data and highlight the similarity and difference in data. The advantage of PCA is that once we have found the pattern we will reduce the number of dimensions (the result that we obtained from PCA retains all the properties of the original dataset).

In this study, we have analyzed PCA, Modified GA and combination of PCA and Modified GA for feature reduction, which helps to determine the most essential features required for classification. Our aim is to find the key features and find which is the best method for feature selection and identify which reduction method have a higher accuracy rate. The Genetic algorithm has been modified by including Modified Keep Best Reproduction Strategy. When using PCA and Modified GA for feature reduction, first the dataset is reduced via PCA, followed by Modified GA. The resultant dataset is then classified.

In our research, we took five datasets, dataset1 (Colon Cancer) from the Bioinformatics group research repository, dataset2 (Breast Cancer Wisconsin (Original) Data Set), dataset3 (Diabetic Retinopathy Debrecen Dataset), dataset4 (Indian Liver Patient Dataset (ILPD)) and dataset5 (Fertility) from the UCI repository. Dataset1 consists of 2001 attributes including class, dataset2, dataset3, dataset4 and dataset5 contains 10,20,10,10 attributes including class respectively. We applied PCA, MGA and a combination of PCA and Modified GA to these datasets. Dataset1 has been reduced to 849 attributes using MGA, 31 attributes using PCA and 31 attributes using combination of PCA and Modified GA. For dataset2**,** 10 attributes have been reduced to 8 attributes using both PCA and Modified GA, and 7 attributes using combined PCA and Modified GA. For dataset3, attributes have been reduced to 9 using PCA, 6 attributes using combination of PCA and Modified GA and 10 attributes using Modified GA. For dataset4, attributes have been reduced to 7 using PCA, 3 using combination of PCA and Modified GA and 6 using MGA. For dataset5, attributes have been reduced to 9 and 10 attributes using PCA and MGA respectively and 6 using combination PCA and Modified GA.

2. DIMENSIONALITY REDUCTION

Dimensionality reduction is the process of reducing the number of attributes using following methods- aggregating, eliminating redundant features, or clustering, for instance. Dimensionality can be reduced by redesigning the features, selecting an appropriate subset among the existing features, and combining existing features. Dimensionality reduction can be divided into two types: feature selection and feature extraction.

The feature reduction process removes redundant or irrelevant features from the original data set. The execution time, classification accuracy and understandability of the feature reduced data set increases and cost of handling of smaller dataset is comparatively low. The irrelevant features can also include noisy data which may have a negative impact in classification accuracy. The feature selection algorithm can be grouped into 3 categories: filters, wrapper and embedded. Wrapper model depends on classification or clustering algorithm, examples of these methods are genetic algorithm, recursive feature elimination algorithm. Filter feature selection algorithm are those which are independent of the classifiers. Embedded models perform feature selection during the learning process [3]. Feature selection for classification can be achieved using association and correlation mechanism [4]. Another approach proposed is to use apriori rule generation algorithm and use it with correlation of attributes to find out closely related attributes [4]. In

feature extraction method the original set of attributes is transformed into a new set of attributes, example of feature extraction is PCA.

2.1. Genetic Algorithm

Genetic algorithm comes under Evolutionary algorithm, GA can be used for a variety of search and optimization problems. Initially GA based learning was used in two different approaches: Pitt approach and Michigan approach. [7]. GA can be used for pattern recognition. Two methods for applying GA for pattern reorganization are,

1) Use GA as a classifier directly in computation
2) Use GA to compute the results.

Another area where GA can be used is for selecting the prototypes in the case-based classification [8]. In GA, solutions to the problem are encoded as chromosomes and a group of chromosomes are known as population. Chromosomes are sets of genes and possible values in the genes are known as alleles. The first generation of chromosome is called the parent generation. The fitness function is applied to the chromosomes to measure the closeness towards the solution. Chromosome with the highest fitness value is used to generate offspring. Offspring can be of different types, parent with best fitness value will automatically survive to the next generation, or two randomly selected parents are taken to generate offspring via different techniques like one point crossover, two point crossover, uniform and nonuniform crossover, etc. or by mutating single parent chromosome. The algorithm stops at when it reaches some threshold. At the end of the last evolution of the algorithm the best chromosome in the population will be the output.

GA uses 3 main types of rules at each step to create the next generation from the current population:

1) Selection rules
2) Crossover rules
3) Mutation rules

Mostly the values of genes will be binary values, but it is not necessary that it should be binary values always. Usage of binary values varies from problems and techniques used in representing chromosomes. For example, binary values in chromosome can be used to represent absence or presence of the fatures; 1 represents presence of the attribute and 0 represents the absence.

Eg. 100011.

In the above binary coded value represents that there are total 7 attributes of which first and the last three attributes is considered to find the solutions.

Darwinian evolution and Natural selection led to a number of models for solution optimization. GA is one of the subsets of this evolution based optimization technique focusing on the application of selection, mutation and recombination of computing problem solution. Since GA is parallel and iterative it is most successfully used in the field like optimization problem, including many pattern recognition and classification task.

Feature reduction using GA which is well-matched to the optimization problem.We have five clinical datasets where Dataset1 consists of 2001 attributes including class, dataset2, dataset3, dataset4 and dataset5 contains 10,20,10,10 attributes including class. Our intension is to reduce the dataset by eliminating unwanted and replicated fields, i.e. if we have given a dataset of n- dimensional input pattern, our task is to use GA to transform the data into m-dimension which is less than n ($m<n$) which maximizes the set of optimization criteria. The transformed data which were reduced using GA are evaluated based on the dimensionality, and either class or correctly classified rate.

2.1.1. GA Based Feature Reduction

In order to compute the feature transform matrix, GA maintains populations to evaluate this matrix, the input patterns multiply by the matrix to produce a set of transforming data which are then sent to the classifier. The samples obtained are divided into a training set and testing set in which the training set is used for training the classifier and the testing set is used to calculate the accuracy. The accuracy rate is passed to the GA in order to measure the quality of the transformation used. GA searches for minimizing the dimensionality of the transformed data by maximizing the accuracy of the classifier.

A direct approach of feature selection was introduced by Siedlecki and sklansky [12]. In this work GA is used to find the optimal binary vector where each bit is referred as an attribute, in this binary 1 refers to the participation of attribute in the classifier and 0 refers to non participation, the resultant feature of the subset is defined using accuracy of the classifier. The GA feature selection has been extended to include binary masking vector along with the feature weight vector of the chromosome. The mask value of 0 defines non participation of attribute for classification, if the value is 1 the field is measured according to the weight

value and include to classifier. The incorporation of mask vector allows the GA to rapidly sample feature with simultaneously optimizing scale factor for feature inclusion.

For each feature a weight value and one or more masking value are assigned. The majority masking value is taken to decide whether the feature is selected or not. Weight vector is used while calculating the fitness value. These vectors are introduced to smooth the GA. When k-NN classifier is used k value is also encoded while encoding chromosome [13].

Advantages of genetic algorithm are:

1) GA can solve optimization problem which can be described with the chromosome encoding
2) GA help solves problems with numerous solution
3) GA is not dependent on the error surface; it helps us to solve multi-dimensional, non-differential, non-continuous, and even non-parametrical problems.
4) Like PCA, GA does not demand the knowledge of mathematics, GA is a method which is very easy to understand.
5) GA is easily reassigned to existing simulations and models.

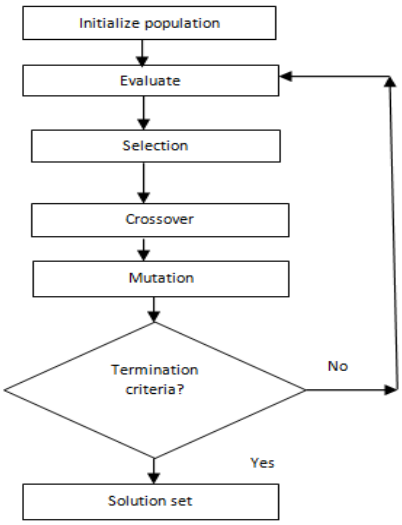

Figure 1. Genetic Algorithm process

2.2. Principal Component Analysis

PCA is a method that is commonly used for feature reduction. It is used for identifying the pattern in the data and highlights the similarity and difference. The main advantage of PCA is that once we have found the pattern we will reduce the number of dimensions of the dataset.

Method:

Step 1: Identify the dataset

We have used clinical cancer dataset which consist of 2000 variable and a class label which decides whether the patient is Normal or has Tumor. Our dataset consist of 63 instances.

Step 2: Subtract the Mean from each Dimension

In the second step, the mean $\bar{x}$ of each dimension is calculated. This mean is then subtracted from each $x_{i;}$ this produces a dataset where the mean is zero.

Step 3: Covariance Matrix Is Calculated

Covariance is measured in multi dimensions. If we calculate the covariance in one-dimension we obtain the variance. Consider a 3-dimensional data set (x, y, z), for which we have to find the cov(x, y), cov(x, z), cov(y, z).

The equation for covariance is given by:

$$\text{Cov(x,y)}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \quad (1)$$

In a Covariance matrix if all the non diagonal elements have positive value it means that X, Y, Z variable increase together.

Step 4: Eigen Vectors and Eigen Values of the Covariance Matrix are calculated.

Here we calculate an Eigen value and Eigen vector from the covariance matrix that we obtained from the previous method.

Step 5: Deriving the new dataset

In the final step of PCA the required Eigen vectors are chosen from the new dataset, obtained from the previous steps. The transpose of the feature vector and new dataset is taken and multiplication is performed.

We have used WEKA for performing PCA First, we load the dataset into weak, then normalize the data using Pre-processing after that we have done the feature reduction the resultant dataset is taken for classification.

Advantages of PCA are:

1) PCA allows us to decouple the feature space
2) PCA is a robust method in image design, data patterns with the help of PCA similarities and differences between them are efficiently identified.
3) Using PCA dimension can be reduced by eliminating redundant information without much loss in the original data.
4) Data can be remodeled and mapped from high dimensional to low dimensional space. The low dimensional space can be resolute using Eigenvectors of the covariance matrix.

3. DATA TRANSFORMATION AND NORMALIZATION

Measurement unit can affect the data analysis. For example, changing the measurement unit of height from inches to meter may lead to very different results. In general, expressing an attribute in smaller unit will lead to a large range for that attribute, and thus tend to give such an attribute greater effect or "weight". To help avoid dependence on the choice of measurement unit the data should be normalized or standardized.

Different Normalization techniques are:

1) Min Max Normalization
2) Nominal to Binary
3) Z-score
4) Decimal Scaling

3.1. Min Max Normalization

Min max normalization performs a linear transformation of data to a new value which fits in the interval [new_min$_A$, new_max$_A$].

$$v_i^{'} = \left(\frac{v_i - min_A}{max_A - min_A}\right) * (new\_max_A - new\_min_A) + new\_min_A \quad (2)$$

3.2. Nominal to Binary

Nominal values are converted into binary values.

3.3. Z Score

Z-score normalization normalizes the values based on the mean and standard deviation. This method is used when the min and max values are unknown or when there are outliers that influence min-max normalization.

$$v_i^{'} = (v_i - \bar{A})/\sigma_A \quad (3)$$

3.4. Decimal Scaling

Decimal scaling normalization normalizes by moving the decimal point of values of attribute A

$$v_i^{'} = v_i/10^j \quad (4)$$

Where j is the smallest integer such that max (|vi'|) <1.

4. PROPOSED SYSTEM

It is known that GA has proven its efficiency in feature reduction. We have augmented two innovations to further improve the efficiency of GA using Modified GA (MGA) and combination of PCA and MGA i.e. Here we combine the best feature of PCA and then combine it with MGA.

4.1. Modified GA (MGA)

In our proposed system we modified the GA by including a Modified Keep Best Reproduction Strategy (MKBR), a midway selection strategy has been used at the end of each generation [14]. MKBR strategy is a modified version of Keep Best Reproduction Strategy (KBR), which overcomes the risk in KBR. In KBR, best offspring from two is selected and is replaced by the best parent. Here there is a risking of losing the better offspring than the next pair of parents. MKBR uses additional selection degrees for determining the survival of parent and offspring.

Modified Genetic Algorithm with Modified Keep-Best Reproduction is given below:

```
Initialize a population of chromosomes;
Evaluate the chromosomes in the population:
While (stopping criteria not reached) do
for i=1 to sizeof(population) do
select 2 chromosomes for recombination;
apply crossover operator to them;
apply mutation operator to them;
evaluate the new chromosomes;
i=i+1;
endfor
compare the parent's fitnesses and remember
the
best parent;
replace the offspring chromosome with lower
fitness by the best parent chromosome;
update stopping criteria;
```

4.2. Combination of PCA and MGA

We propose a system in which the dataset is first preprocessed using a normalization method; to this we then apply PCA. After applying PCA, to the resultant dataset MGA is applied to increase the accuracy.

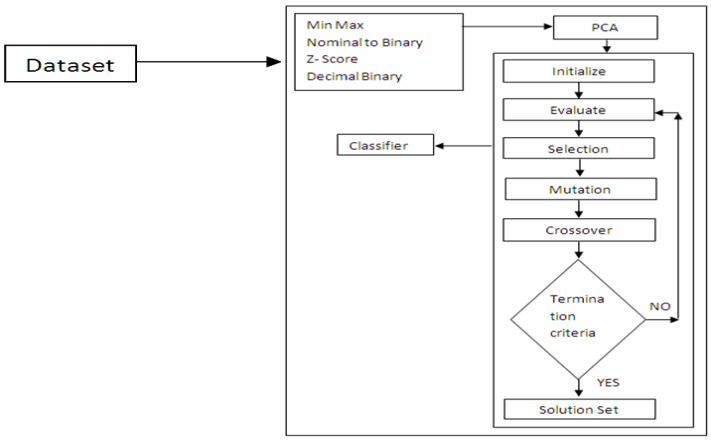

Figure 2. Combination of PCA and MGA for feature reduction

The dataset is first normalized. This is done to remove noisy data present in the dataset. Then PCA a dimensionality reduction technique is applied to the normalized dataset. PCA reduces the number of attributes to a fewer number of attributes. The modified Genetic algorithm is applied to the resultant dataset. MGA selects only the attributes that satisfy the fitness function resulting into a subset of the dataset. These subsets are then classified and compute the correctly classified rate.

5. RESULTS AND ANALYSIS

We have used three types of feature reduction methods PCA, combination of PCA and MGA, and Modified genetic algorithm and the resulted dataset are taken for classification. Feature reduction is done in order to increase the accuracy of the data sets. Before feature reduction the accuracy of dataset1, dataset2, dataset3, dataset4 and dataset5 were 82.2% , 53.2%, 56%, 55.7% and 85%.

5.1. Result

We have obtained a classification accuracy of 72.5%, 94.1%, 58%, 57.5% and 86% for the datasets using PCA. Datasets yielded a classification accuracy of 74.1%, 96.5%, 59%, 70% and 85% for MGA. An improved classification accuracy has been obtained using combination of PCA and MGA for all the datasets. The improved accuracy for the datasets are 83.2%, 97.2%, 66%, 71% and 88%.

Table 1. accuracy using various features reduction method

	Classification accuracy, using Feature reduction		
	PCA	**PCA+MGA**	**MGA**
Dataset1	72.5%	83.2%	74.1%
Dataset2	94.1%	97.2%	96.5%
Dataset3	58%	66%	59%
Dataset4	57.5%	71%	70%
Dataset5	86%	88%	85%

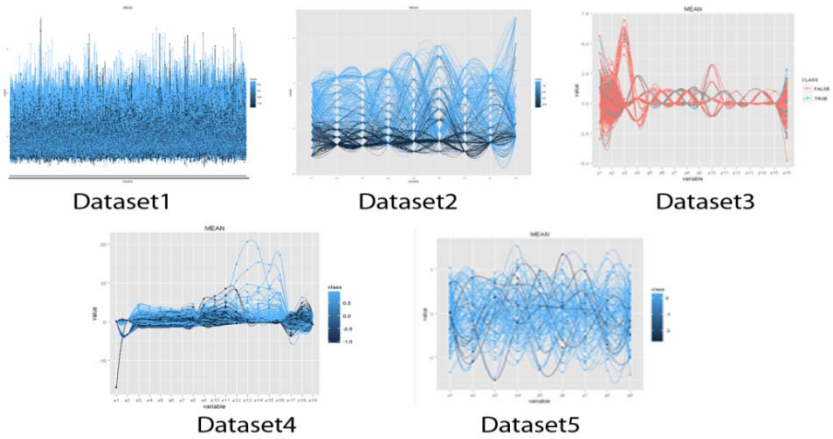

Figure 3. Graphical representation of classified datasets without feature reduction

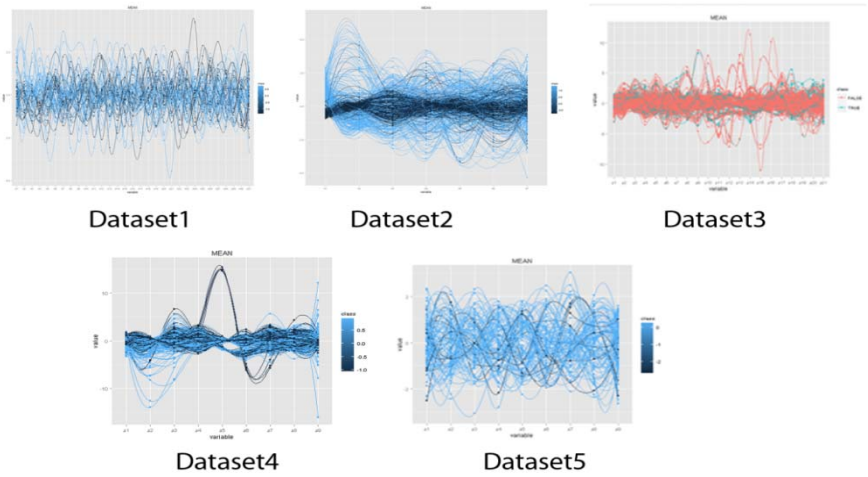

Figure 4. Graphical representation of classified datasset using PCA

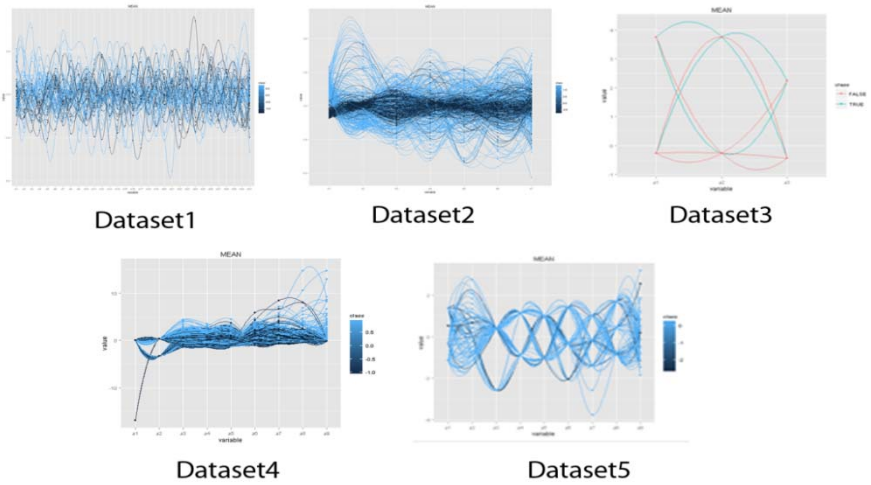

Figure 5. Graphical representation of classified dataset using MGA

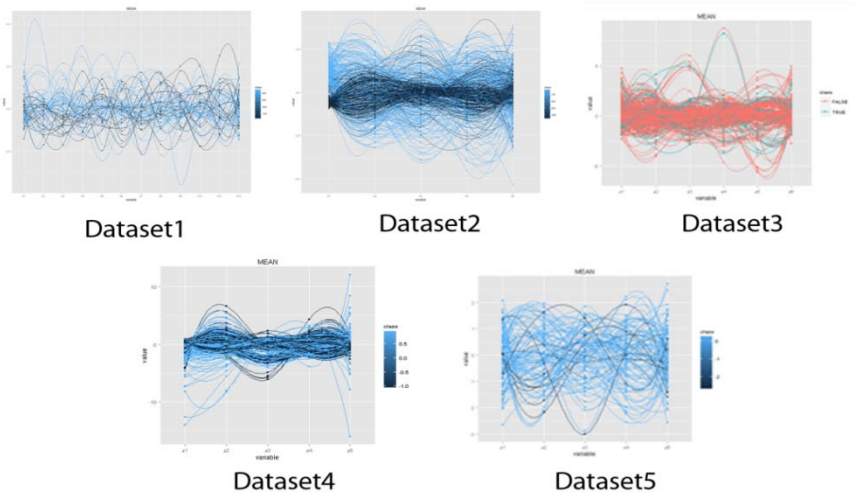

Figure 6. Graphical representation of classified dataset using combination of PCA and Modified GA

From the experiment we conclude that modified GA has a higher accuracy rate compared to other feature reduction methods.

6. CONCLUSION

In this paper, we use three different types of feature reduction methods, namely PCA, MGA and combination of PCA and MGA to identify the key factor for the diagnosis of the diseases. The above mentioned feature reduction methods are applied to the datasets, and the accuracy has been computed. The correctly classified rate using PCA on datasets are 72.5%, 94.1%, 58%, 57.5% and 86%. Datasets yielded a classification accuracy of 74.1%, 96.5%, 59%, 70% and 85% for MGA. An improved classification accuracy has been obtained using combination of PCA and MGA for all datasets. Datasets showed an increased accuracy of 83.2%, 97.2%, 66%, 71% and 88%. From the results, we conclude that combination of PCA and MGA have a higher accuracy rate compared to others.

ACKNOWLEDGEMENTS

This work is supported by the DST Funded Project, (SR/CSI/81/2011) under Cognitive Science Research Initiative in the Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham University, Kochi.

REFERENCES

[1] Yohannes Kassahun, Roberta Perrone, Elena De Momi, Eimar Berghofer, Laura Tassi, Maria Paola Canevini, Roberto Spreafico, Giancarlo Ferrigno, and Frank Kirchner. "Automatic classification of epilepsy types using ontology-based and genetic-based machine learning", *Artificial Intelligence in Medicine.* Vol. 61, No. 2, pp. 79-88, 2014.

[2] Hongmei Yan, Jun Zheng,Yingtao Jiang, Chenglin Peng, Shozhoug Xia. "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm", *Applied soft computing,* Vol. 8, No. 2, pp. 1105-1111, 2008.

[3] Esra Mahsereci Karabulut, Selma Ayse Ozel, and Turgay Ibrikci. "A Comparative Study on the effect of feature selection on classification accuracy", Procedia Technology.Vol. 1, pp. 323-327, 2012.

[4] Prof. K Rajeswari, Dr. v.vaithiyanathan, and Shailaja V. Pede. "Feature Selection for Classification in Medical Data Mining", *International Journal of emerging trends and technology in computer science,* Vol. 2, No. 2, pp. 492-497, 2013.

[5] Jin-Xing Hao, Yan Yu, Rob Law, and Davis Ka Chio Fong. "A genetic algorithm-based learning approach to understand customer satisfaction with OTA websites", *Tourism Management,* Vol. 46, pp. 231-241, 2015.

[6] Abdulhamit Subasi, and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines", *Expert Systems with Applications,* Vol. 37, No. 12, pp. 8659-8666, 2010.

[7] Ian W. Flockhart, and Nicholas J. Radcliffe, "A genetic algorithm- based approach to data mining", *KDD-96 Proceedings,* pp. 299-302, 1996.

[8] Gunjan Verma, and Vineeta Verma, "Role and application of genetic algorithm in data mining", *International journal of computer applications,* Vol. 48, No. 17, pp. 5-8, 2012.

[9] Xcechuan Wang, Kuldip K. paliwal, "Feature Extraction and Dimensionality Reduction Algorithm and their Application in Vowel Recognition", *Pattern Recognition,* Vol. 36, No. 10. Pp. 2429-2439, 2003.

[10] Rashedur M. Rahman, and Fazle Rabbi Md. Hasan, "Using and Comparing different decision tree classification technique for datamining ICDDR,B Hospital Survellance data", *Expert System with Application*, Vol. 38, No. 9, pp. 11421-11436, 2011.

[11] D. Nithya, V. Suganya, and R. Saranya Irudayan Mary, "Feature Selection using Integer and Binary Coded Genetic Algorithm to improve the performance of SVM classifer", *Journal of Compute Application,* Vol. 6, No. 3, pp. 57-61, 2013.

[12] W. Siedlecki and J. Sklansky, "A note on genetic algorithm for large scale feature selection", *Pattern Recognit Letters,* Vol. 10, No. 5, pp. 335-347, 1989.

[13] Michael L. Raymer, William F. Punch, Erik D. Goodman, Leslie A. Kuhn, and Anil K. Jain. "Dimensionality Reduction using Genetic algorithms". *IEEE trasaction on evolutionary computation,* Vol. 4, No. 2, pp. 164-171, 2000.

[14] M. V. Judy, K. S Ravichandran. "A Solution to protein folding problem using a Genetic Algorithm with modified keep best reproduction strategy", *Evolutionary Computation,* pp. 4776-4780, 2007.

BIOGRAPHIES OF AUTHORS

Srividya Sivasankar, Post Graduated in Master of Computer Applications from Amrita Vishwa vidyapeetham, in 2015. Her area of interest includes Data mining.

Sruthi Nair, Post Graduated in Master of Computer Applications from Amrita Vishwa vidyapeetham, in 2015. Her area of interest includes Programming.

Dr M V Judy, PhD in Computer Science is an Associate Professor and Head of the Department of CS and IT at Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi. She is the Principle Investigator of a project under Department of Science and Technology (DST), Government of India. Her research interests include computational biology, machine learning and data analytics.