An Implementation of genetic algorithm based feature selection approach over medical datasets

Dr. A. Shaik Abdul Khadir[#1], K. Mohamed Amanullah[#2]
[#1]Research Department of Computer Science, KhadirMohideen College, Adirampattinam, India
[1]Mail id: asak_cs_kmc@ymail.com
[#2]Research Department of Computer Science, Bishop Heber College, Tirchy, India
[2]Mail id: kmabhc@gmail.com

***Abstract*-One of the heuristic approaches that can be applied to many real world applications for the attainment of optimized solutions is the Genetic Algorithm (GA). Feature selection techniques of data mining alsotake the advantages of genetic algorithm for extracting the meaningful attributes from high dimensional datasets. This paper presents an improved genetic algorithm for the same feature selection process for the enhancement of classification or clustering results. A Multiple Linear Regression (MLR) technique is employed as fitness function to identify the best influencing attribute for building a knowledge prediction model. The results of the MLR-GA have outperformed the existing feature selection algorithms in terms of accuracy.**

Keyword - Genetic Algorithm, Feature Selection, MLR, Fitness Function, High Dimensional

I. INTRODUCTION

Data mining comprises of techniques that analyzes the meaningful facts that are hidden in large data stores. The tasks of data mining can be organized into data collection, cleaning, data extraction and predictive or descriptive analysis of data with which the extraction of meaningful data is the most important task for enhancing the outcome of facts.

Data extraction also called feature selection can be viewed as a search technique that proposes new subset of features through an evaluation measure based on ranking the attributes. The ranking of the attributes are computed by selecting the subset of features that minimizes the error rate. Feature selection performs a complete search over the space for identifying the best attributes which may increase the computational cost for all but the smallest of feature sets. Despite the types of feature selection algorithm are majorly classified into filter and wrapper methods, now-a-days the evolutionary algorithms have also implemented to obtain the optimized features.

The evolutionary algorithms are meta-heuristics algorithms that employ population based optimized solutions in search space. Genetic algorithm stands first and best ever known evolutionary algorithm that imitates the natural process of gene duplication such as selection, cross-over and mutation for the reproduction of chromosomes. The algorithm is initialized with few random individuals called initial populations (IPs) which are initially considered sub optimized solutions which can be refined in the consecutive iterations. The fitness of IPs is calculated next, for assessing their ability in the prediction task. During each iteration, a set of best individuals are chosen to breed a new offspring through cross-over and mutation operations and the fitness of the new offspring is calculated to replace the least fitness valued individuals. In this paper, a novel genetic algorithm is proposed to calculate the fitness of the attributes to extract the best features that most contribute to the prediction task.

II. REVIEW OF LITERATURE

Abualigah et al. [1] have proposed a genetic algorithm based unsupervised feature selection method (FSGATC) for extracting the best subset of features that obtains accurate clusters. The authors have used mean absolute difference method as fitness function, uniform two point as mutation operator and a probability based parameter as cross over operator. The authors have claimed that the proposed FSGATC have improved the performance of text clustering with highest accuracy and suggested to collaborate the proposed work with other meta-heuristics algorithms to improve global search to find more accurate clusters.

Kashyap et al. [2] have proposed a multi objective genetic algorithm for feature selection with the objective of maximizing the Laplacian score which aims at analyzing the importance or relevance of features and minimizing the inter-attribute correlation which aims at analyzing the dependency of feature. The authors have set the mutation, cross over and Laplacian score probability as 0.7, 0.05 and 0.5 respectively to perform the multi objective GA operations and claimed that their method has achieved feature set reduction by removing

62.27% noisy data or removed. The authors have stated that tuning level of parameters should be enhanced to measure the importance of features and inter feature dependency.

Bian et al. [3] have developed a cost-sensitive feature selection algorithm called CSFSG that adds cost-based evaluation function of filter feature selection using a chaos genetic algorithm. Their method evaluates both feature acquiring costs (test costs) and misclassification costs in network security domain through which the influence weak instances are discarded from majority of classes. The authors have claimed that their work is tested on a large scale dataset in network security and have stated that the proposed work have improved the classification accuracy with decreased classification time only with low-dimensional datasets.

Singh et al. [4] have proposed four types of genetic cross over methods such as AND, XOR, OR and Single Point to be used to generate new chromosomes. The authors have executed the reduced features with the several classification algorithms and have proven that their methods achieve better accuracy than the existing feature selection algorithms.

Desale et al. [5] have proposed a mathematical intersection principle based approach using genetic algorithm with correlated attributes for feature selection. The authors have tested the effect of the proposed work over different datasets with two classification algorithms such as Naïve Bayes and J48 and have stated that the method is able to select minimum number of features from the dataset having numerical data type. But, the method works only for numerical data types.

III. METHODOLOGY

The methodology of the MLRGA consists of the series of tasks that contains the essential operations of genetic algorithm. In the proposed work, an individual chromosome is represented by a vector of attribute length (M) that consists of the values of any instance i, rather than considering a bit string of length M consists of 0's or 1's as it is done in many GA implementations. A population is a group of individual chromosomes represented by a matrix of dimension of population size (N) and M. The rows of a matrix C1…Cn represents the values that the chromosomes holds for the attributes a1.. an. The columns of the matrix denote the length of attributes in the dataset. The evaluation of the subset of variables is estimated through a Multiple Linear Regressive fitness function which assigns the correlation coefficients of attributes as weights for the values of the individual chromosomes. After identifying the fitness of each chromosome, the algorithm selects two random individuals as parents and applies the genetic operators such as cross over and mutation to produce new off springs for describing the best subset of attributes. The derivation of subset of attributes through MLRGA is described as the following:

A. Fitness Function

Fitness function is an important objective function to evaluate the optimality of each chromosome. For each chromosome in the population matrix the fitness function is computed through multiple linear regression function. The multiple linear regression is the most common form of linear regression analysis which is used to explain the relationship between one continuous dependent variable from two or more independent variables by fitting a linear equation, which is denoted using the notion presented in Equation 1 [6].

$$\beta_y = \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \cdots + \mu_1 x_1 + \mu_p x_p \tag{1}$$

The line describes how mean the response β_y change with the explanatory variables $x_1 \ldots x_p$ and assumed to have the same standard deviation σ [7]. Finally, a solution with high fitness value is considered as the optimal solution.

B. Cross over

Cross over generates the new off springs from two parent chromosomes by swapping the selected variables from each parent. Among the four single, two, uniform and arithmetic cross overs, the proposed work employs the uniform cross over method for randomly copying values from parent 1 to parent 2 chromosomes. Table 1 denotes the cross over operation of uniform random cross over operation of sample bit strings.

Parent 1	**0**	1	1	**1**	**0**	**1**	0	0
Parent 2	1	**0**	**1**	1	1	1	**1**	**0**
Child 1	0	0	1	1	0	1	1	0

C. Mutation

Mutation operator is intended to derive new off springs from a single parent chromosome by flipping the values of the parent by predetermined number of times for generating better chromosomes. In this work, the uniform random mutation flip is used as mutation operator. Table 2 denotes the mutation operation of uniform random mutation of sample bit string.

Parent 1	**0**	1	1	**1**	**0**	1	0	0
Child 1	1	1	1	0	1	1	1	0

The results of the fitness values are is stored in a vector 'v' to find out the best optimal solution for evaluating the best subset of attributes in the dataset. The operators of fitness calculation, cross over and mutation are executed iteratively until the optimized solution is reached or the maximum number of generation is reached. In this paper, the maximum number of iteration is set to K=200. Figure 1 depicts the pictorial representation of the MLRGA.

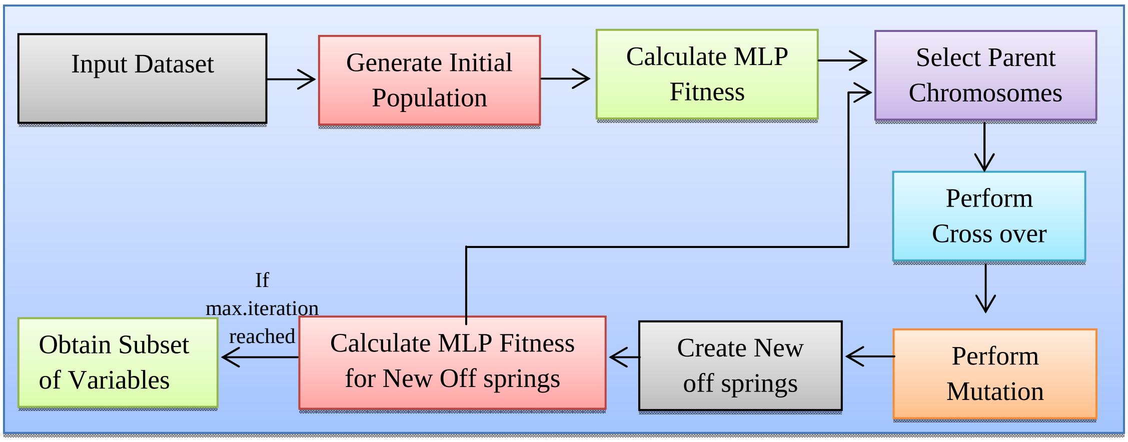

Input Dataset
Generate Initial Population
Calculate MLP Fitness
Select Parent Chromosomes
Perform Cross over
Perform Mutation
Create New off springs
Calculate MLP Fitness for New Off springs
If max.iteration reached
Obtain Subset of Variables

Fig.1. MLRGA Framework

IV. EXPERIMENTATIONAND RESULT DISCUSSIONS

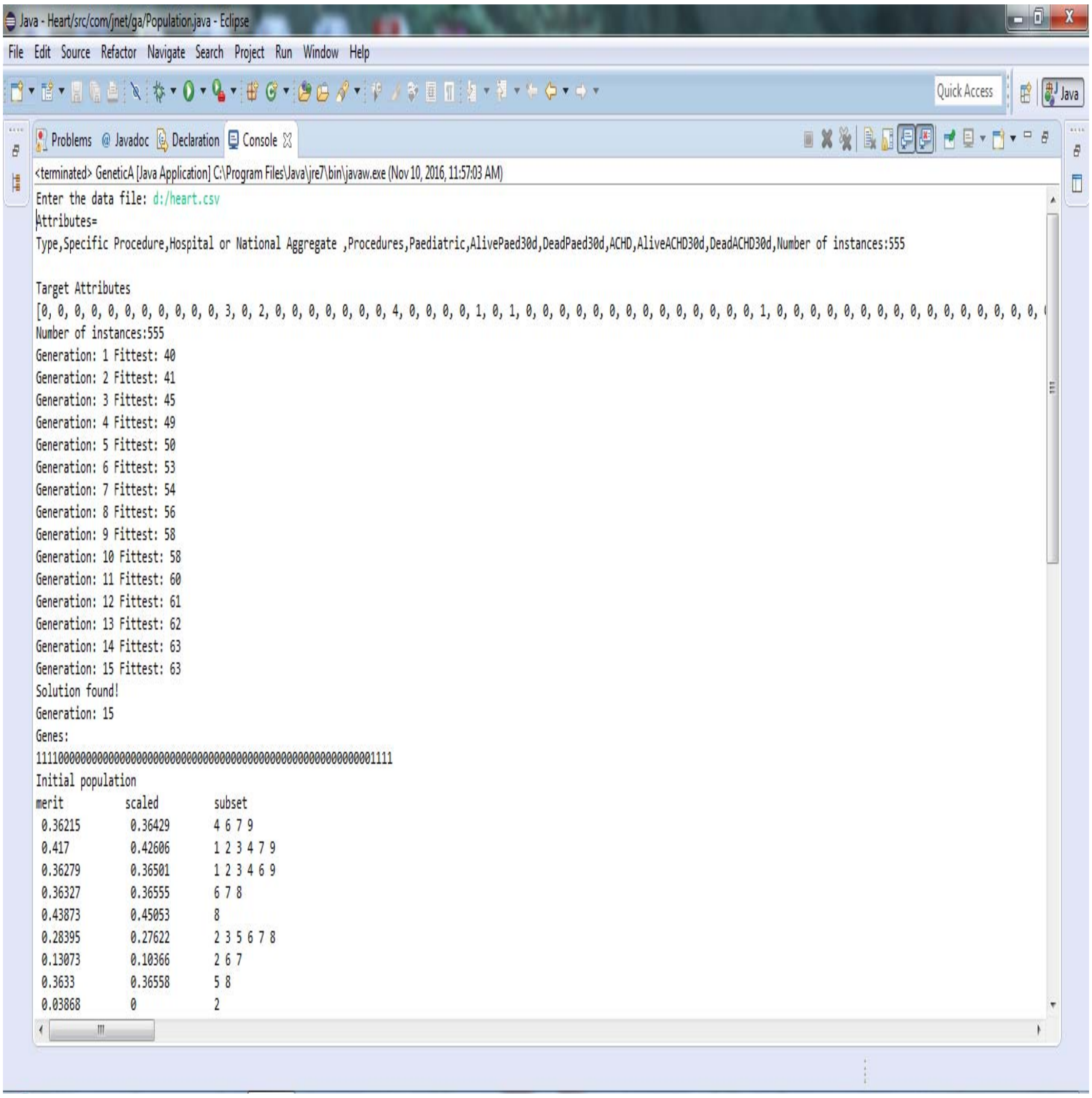
For the experimentation, a heart disease dataset which is extracted from Congenital Heart Disease (CHD) datasets is used in this study. The dataset originally contains 10 attributes including the predictor with 700 instances. The dataset contains 30 day outcomes for congenital treatment in England, although the audit covers all of the UK and the republic Ireland [8]. The objective of the experiment is to extract the subset of attributes so as to improve the prediction accuracy of heart disease. The description of the dataset is listed in Table 1.

TABLE 1. Description of Heart Disease Dataset

Attribute Name	Description	Data Type
Specific Procedure	Procedure for treatment of congenital heart disease	Text
Hospital	Congenital heart disease	Text
Procedures	Total Number of procedures	Text
Pediatric	Total number of pediatric procedures	Integer
AlivePaed30d	Total number of pediatric procedures (age<16) - patient alive at 30 days after procedure	Integer
DeadPaed30d	Total number of paediatric procedures (age <16) - patient dead at 30 days after procedure	Integer
ACHD	Total number of Adult Congenital Heart Disease (ACHD) procedures (age <16)	Integer
AliveACHD30d	Total number of ACHD procedures (age <16) - patient alive at 30 days after procedure	Integer
DeadACHD30d	Total number of ACHD procedures (age <16) - patient dead at 30 days after procedure	Integer
Class	Existence of the disease	Binary

Two different types of procedures are implemented for constructing the training set classifier for heart disease dataset. The first procedure is designed to execute the classifier over the original dataset by evading the preprocessing steps for being compared with our proposed work. The second procedure is designed to execute the classifier over the subset of attributes containing the selected features resulted from MLRGA algorithm. This will result in analyzing the performance of the MLRGA over the classification performance. The dataset is divided into training and test sets, and 10-fold cross validation is used to segregate the sets. Thus, 90% of instances in heart disease dataset is used for training the MLRGA model and 10% of the dataset is used for testing the model.

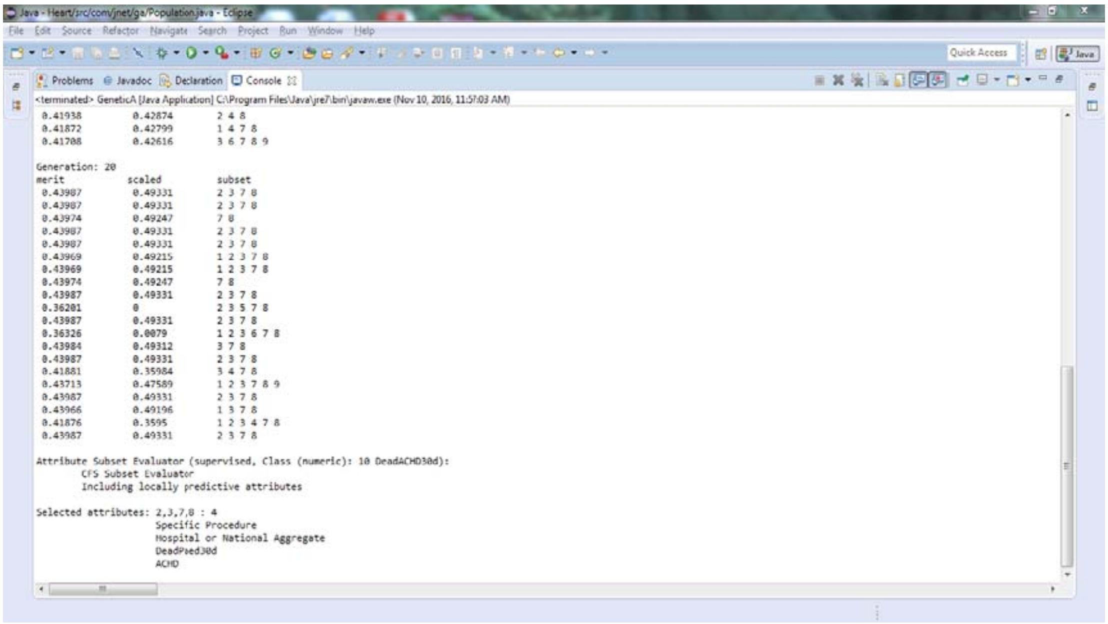
The proposed MLRGA algorithm is developed as a software program in JAVA. The code consists of all possible operators in GA as specified in the methodology. The fitness function of the software is implemented with multiple linear regression method for analyzing the usefulness of individual attributes. The initial population size of the training data is set to 100 and the cross over and mutation rates are set as 0.9 and 0.001 respectively. Figure 2 denotes the execution of MLRGA algorithm over heart disease dataset.


Java - Heart/src/com/jnet/ga/Population.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Problems @ Javadoc Declaration Console

```
<terminated> GeneticA [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Nov 10, 2016, 11:57:03 AM)
Enter the data file: d:/heart.csv
Attributes=
Type,Specific Procedure,Hospital or National Aggregate ,Procedures,Paediatric,AlivePaed30d,DeadPaed30d,ACHD,AliveACHD30d,DeadACHD30d,Number of instances:555

Target Attributes
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 2, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
Number of instances:555
Generation: 1 Fittest: 40
Generation: 2 Fittest: 41
Generation: 3 Fittest: 45
Generation: 4 Fittest: 49
Generation: 5 Fittest: 50
Generation: 6 Fittest: 53
Generation: 7 Fittest: 54
Generation: 8 Fittest: 56
Generation: 9 Fittest: 58
Generation: 10 Fittest: 58
Generation: 11 Fittest: 60
Generation: 12 Fittest: 61
Generation: 13 Fittest: 62
Generation: 14 Fittest: 63
Generation: 15 Fittest: 63
Solution found!
Generation: 15
Genes:
1111000000000000000000000000000000000000000000000000000000001111
Initial population
merit          scaled          subset
 0.36215        0.36429        4 6 7 9
 0.417          0.42606        1 2 3 4 7 9
 0.36279        0.36501        1 2 3 4 6 9
 0.36327        0.36555        6 7 8
 0.43873        0.45053        8
 0.28395        0.27622        2 3 5 6 7 8
 0.13073        0.10366        2 6 7
 0.3633         0.36558        5 8
 0.03868        0              2
```

Fig. 2. Execution of MLRGA

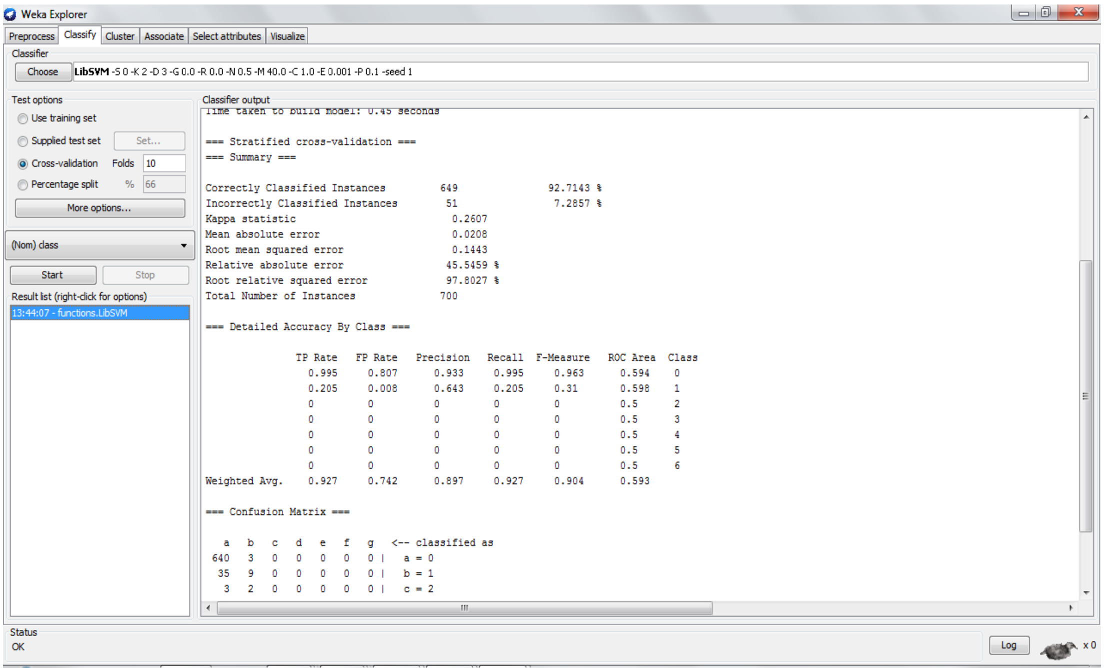
The results obtained by the MLRGA with the reduced subset of attributes of the experimented dataset are depicted in Figure 3. The best subset of attributes that are extracted from the heart disease through the proposed MLRGA is specific procedure, National Aggregate, DeadACHD30d, ACHD.


Java - Heart/src/com/jnet/ga/Population.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Problems
Javadoc
Declaration
Console

```
<terminated> GeneticA [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Nov 10, 2016, 11:57:03 AM)
 0.41938         0.42874         2 4 8
 0.41872         0.42799         1 4 7 8
 0.41708         0.42616         3 6 7 8 9

Generation: 20
merit           scaled          subset
 0.43987         0.49331         2 3 7 8
 0.43987         0.49331         2 3 7 8
 0.43974         0.49247         7 8
 0.43987         0.49331         2 3 7 8
 0.43987         0.49331         2 3 7 8
 0.43969         0.49215         1 2 3 7 8
 0.43969         0.49215         1 2 3 7 8
 0.43974         0.49247         7 8
 0.43987         0.49331         2 3 7 8
 0.36201         0               2 3 5 7 8
 0.43987         0.49331         2 3 7 8
 0.36326         0.0079          1 2 3 6 7 8
 0.43984         0.49312         3 7 8
 0.43987         0.49331         2 3 7 8
 0.41881         0.35984         3 4 7 8
 0.43713         0.47589         1 2 3 7 8 9
 0.43987         0.49331         2 3 7 8
 0.43966         0.49196         1 3 7 8
 0.41876         0.3595          1 2 3 4 7 8
 0.43987         0.49331         2 3 7 8

Attribute Subset Evaluator (supervised, Class (numeric): 10 DeadACHD30d):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,7,8 : 4
                     Specific Procedure
                     Hospital or National Aggregate
                     DeadPsed30d
                     ACHD
```

Fig. 3. MLRGA Feature Extraction

The reduced attributes are then inputted WEKA data mining tool to be executed onto support vector machine classifier to build a knowledge prediction model. The user should ensure the class path file for running support vector machine classifier in WEKA tool. Figure 4 denotes performance of SVM classifier with reduced dataset.



Classifier
LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)
13:44:07 - functions.LibSVM

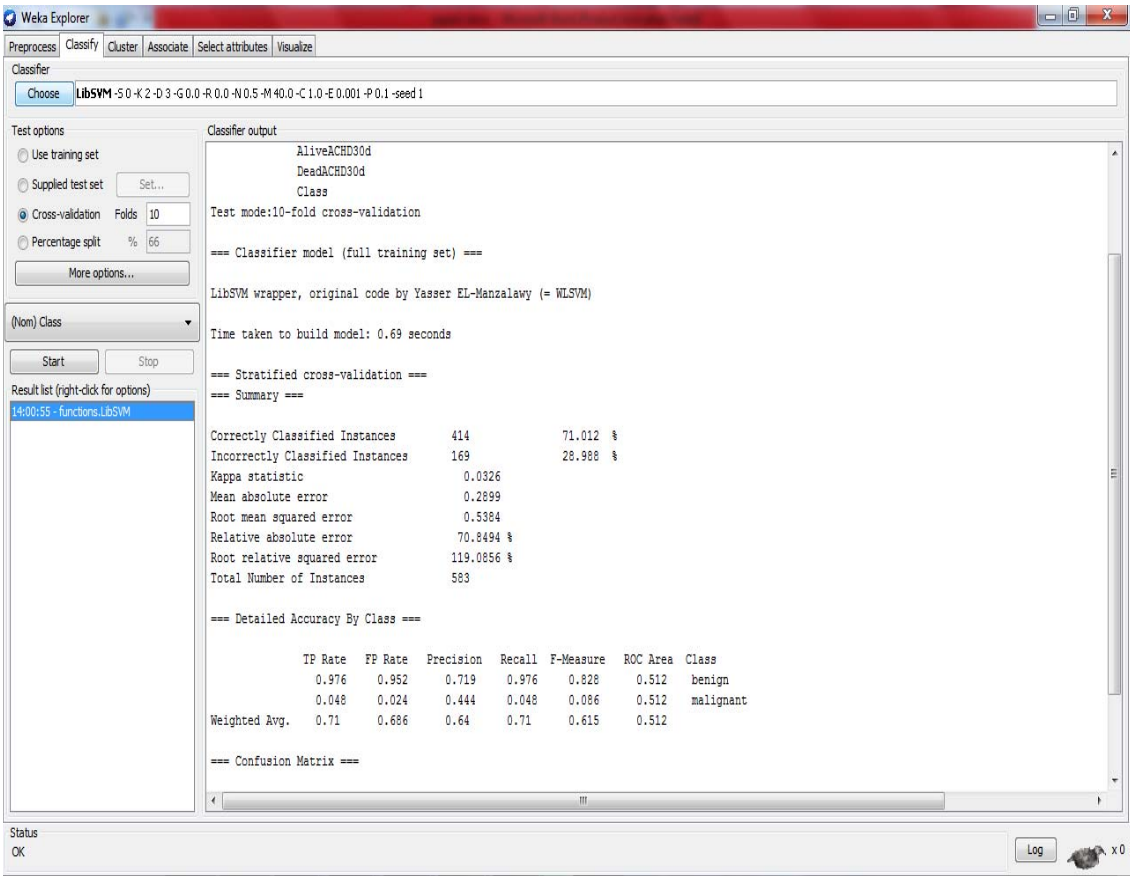
Classifier output

```
Time taken to build model: 0.45 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         649               92.7143 %
Incorrectly Classified Instances        51                7.2857 %
Kappa statistic                          0.2607
Mean absolute error                      0.0208
Root mean squared error                  0.1443
Relative absolute error                 45.5459 %
Root relative squared error             97.8027 %
Total Number of Instances              700

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995     0.807     0.933      0.995     0.963      0.594    0
                 0.205     0.008     0.643      0.205     0.31       0.598    1
                 0         0         0          0         0          0.5      2
                 0         0         0          0         0          0.5      3
                 0         0         0          0         0          0.5      4
                 0         0         0          0         0          0.5      5
                 0         0         0          0         0          0.5      6
Weighted Avg.    0.927     0.742     0.897      0.927     0.904      0.593

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 640   3   0   0   0   0   0 |   a = 0
  35   9   0   0   0   0   0 |   b = 1
   3   2   0   0   0   0   0 |   c = 2
```

Status
OK

Fig. 4. Execution MLRGA with SVM

Figure 5 denotes the performance of svm classifier with original dataset. the accuracy obtained by the instances that are correctly identified with original dataset is 71.012%. Moreover, the time taken to build the knowledge prediction model for the original dataset is 0.69 seconds.



Classifier: LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

14:00:55 - functions.LibSVM

Classifier output

```
              AliveACHD30d
              DeadACHD30d
              Class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.69 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       414               71.012  %
Incorrectly Classified Instances     169               28.988  %
Kappa statistic                        0.0326
Mean absolute error                    0.2899
Root mean squared error                0.5384
Relative absolute error               70.8494 %
Root relative squared error          119.0856 %
Total Number of Instances            583

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0.952      0.719     0.976     0.828      0.512    benign
                 0.048     0.024      0.444     0.048     0.086      0.512    malignant
Weighted Avg.    0.71      0.686      0.64      0.71      0.615      0.512

=== Confusion Matrix ===
```

Status

OK

Fig. 5. Execution of MLRGA with Original Dataset

The comparison of MLRGA dataset with the original dataset is compared with time and accuracy and shown in table

TABLE 2. Performance Analysis of MLRGA

Dataset	Time (in seconds)	Accuracy
Congenital MLRGA dataset	0.45 seconds	92.7143%
Congenital Original dataset	0.69 seconds	72.012%

Figure 6 denotes the time comparison of congenital original dataset with reduced MLRGA dataset which shows the time taken by the proposed work is minimum than the original dataset with 0.45 seconds.

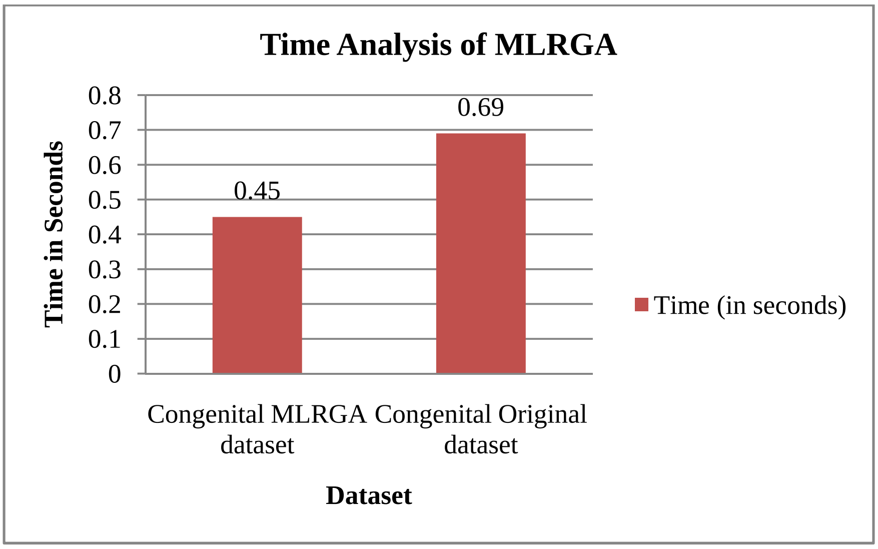


Fig. 6. MLRGA Time Analysis

Figure 7 denotes the accuracy analysis of reduced MLPGA Vs Support Vector Machine dataset with the original congenital dataset Vs Support Vector Machine. The accuracy attained by the proposed work is higher than the original dataset with 92.7143% accuracy.

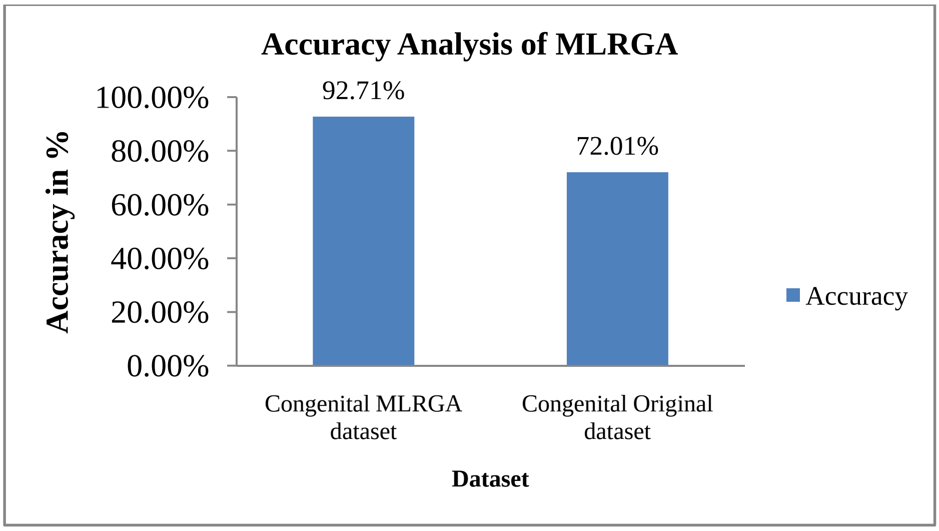


Fig. 7. MLRGA Accuracy Analysis

V. Conclusion

Feature selection is an important process in data mining. The main objective of feature selection step is to derive a qualitative predictive or descriptive model by removing the noisy or inconsistent data to improve the accuracy of the knowledge prediction model. The current research of feature selection focuses on the evolutionary based attribute selection so as to increase the accuracy of prediction. In this paper, an MLRGAmethod that applies the multiple linear regressive as fitness function in genetic algorithm is proposed for extracting the meaningful attributes through meta-heuristics optimization approach. The features that are derived by the MLRGA are executed upon the principal component analysis classifier method to evaluate the prediction accuracy with reduced set. The results have proven that the accuracy obtained by the MLRGA with SVM is greater than the original dataset with SVM. Moreover, the research findings of the paper also provide several guidelines for implementing suitable procedures for performing feature selection. In future, this work can further be extended to other evolutionary algorithms to optimize feature selection process.

References

[1] Abualigah, Laith Mohammad, AhamadTajudinKhader, and Mohammed Azmi Al-Betar. "Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering." In Computer Science and Information Technology (CSIT), 2016 7th International Conference on, pp. 1-6. IEEE, 2016.

[2] Kashyap, Himanshu, Sohini Das, JayeeBhattacharjee, RituHalder, and SaptarsiGoswami. "Multi-objective Genetic Algorithm setup for feature subset selection in clustering." In Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on, pp. 243-247. IEEE, 2016.

[3] Bian, Jing, Xin-guangPeng, Ying Wang, and Hai Zhang. "An Efficient Cost-Sensitive Feature Selection Using Chaos Genetic Algorithm for Class Imbalance Problem." Mathematical Problems in Engineering 2016 (2016).

[4] Singh, D. Asir Antony Gnana, E. JebamalarLeavline, R. Priyanka, and P. Padma Priya. "Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis." International Journal of Intelligent Systems and Applications 8, no. 1 (2016): 67.

[5] Desale, Ketan Sanjay, Balaji Mane, PrashantBerkile, and SushantShivale. "Effective Feature Selection Approach using Genetic Algorithm for Numerical Data."

[6] Christy, Joy, and S. Hari Ganesh. "Linear Regressive Clustering."International Journal of Advanced Engineering and Global Technology, Vol-04, Issue-04, July 2016.

[7] Ari, Bertan, and H. Altay Güvenir. Clustered linear regression. Knowledge-Based Systems 15, no. 3 (2002), pp: 169-175

[8] https://data.gov.uk/dataset/congenitalheartdisease

DOI: 10.21817/ijet/2017/v9i3/170903515