Feature Selection using Integer and Binary coded Genetic Algorithm to improve the performance of SVM Classifier

D.Nithya [a,*], V.Suganya [b,1], R.Saranya Irudaya Mary [c,1]

Abstract - **This paper presents, a Feature Selection using Integer and Binary coded Genetic Algorithm to improve the performance of SVM Classifier. Data Mining (DM) is the process of exploration and analysis, by automatic or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules. DM methods can be divided into supervised and unsupervised learning techniques. Classification is a supervised learning technique. In this paper classification algorithms like Support Vector Machine (SVM) and Genetic Algorithm (GA) are used to find the classification accuracy for the Wisconsin Breast Cancer dataset. SVM is a class of supervised learning method. In SVM, Radial basis function and Polynomial kernel function are used. GA is a search and optimization method. In GA, Integer and Binary Coded Genetic Algorithm are used. Feature Selection is used to improve the accuracy of the SVM classifier. The comparison of Support Vector Machine and Genetic Algorithm are performed based on the classification accuracy and run time.**

Index Terms – **Data Mining, SVM, Genetic Algorithm, Feature Selection.**

I. INTRODUCTION

Data Mining (sometimes called data or knowledge Discovery Data (KDD)) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization and information science (Figure 1) [2]. DM methods can be divided into supervised and unsupervised learning techniques. Supervised learning techniques are: Classification, Regression and Attribute Importance. Unsupervised learning techniques are: Clustering, Association and Feature Extraction.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. A classification SVM model attempts to separate the target classes with the widest possible margin.

Manuscript received 05/September/2013.

D.Nithya, Assistant Professor, Department of Computer Science and Engineering, Avinashilingam University, Coimbatore, India, (E-mail: nittudeva@gamil.com).

V.Suganya, Assistant Professor, Department of Computer Science and Engineering, Avinashilingam University, Coimbatore, India, (E-mail: suganyadhayanesh@gamil.com).

R.Saranya Irudaya Mary, Assistant Professor, Department of Computer Science and Engineering, Avinashilingam University, Coimbatore, India,(E-mail: maryregis26@gamil.com).

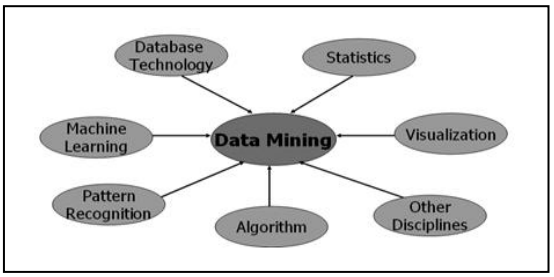

Figure 1. Classification of Data Mining

A regression SVM model tries to find a continuous function such that maximum number of data points lie within an epsilon-wide tube around it. SVM Classification is an active research area which solves classification problems in different domain. The performance of SVM largely depends on the kernel functions. Basically there are 4 types of kernel functions. They are linear kernel functions, Polynomial kernel functions, Radial basis kernel functions and sigmoid kernel functions.

Genetic Algorithm (GA) is general purpose of search and optimization technique [3]. GA is a search technique used in computing to find exact or approximate solutions to optimization and search problems. GA proceeds to initialize a population of solutions randomly, and then improve it through repetitive application of mutation, crossover, inversion and selection operators. In this paper, GA depends on Integer and Binary Coded Genetic Algorithm for improving the accuracy.

Feature selection (FS), also known as variable selection, feature reduction, attribute selection or variable subset selection. Feature selection has been effective in removing irrelevant and redundant features, increasing efficiency in mining tasks, improving mining performance like predictive accuracy, and enhancing result comprehensibility [5].

The benchmark dataset, Wisconsin Breast Cancer Dataset is obtained from the UCI (University of California at Irvine) Machine Learning Repository which was donated by Dr. William H. Wolberg [6]. The dataset consists of 699 instances divided into 2 classes namely Benign and Malignant, each with 11 attributes.

The organization of this paper is as follows. In Section II, methodology used is elaborated. In Section III, the experimental results are discussed and finally section IV provides the Conclusion and Future work.

II. METHODOLOGY

A. Dataset Description

The Wisconsin Breast Cancer Dataset contains 11 numeric attributes and 699 instances. The classes include integers valued 2 (benign) and 4 (malignant). It consists of linear and nominal attributes. The data set consists of 11 numeric attributes which include Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion Single, Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class respectively. It has 699 instances of which 458 are benign and 241 are malignant.There are 16 missing values in the attribute Bare Nuclei. The description of the dataset and missing value attribute are tabulated in I and II.

Table I. Dataset Description

Dataset	Wisconsin Breast Cancer
No. of Attributes	11
No. of Instances	699
No. of Classes	2 (2-Benign, 4-Malignant)

Table II. Missing Attribute

S.No.	Name of the Attribute	Missing Values	Class Attributes	
			2	4
1.	Bare Nuclei	16	14	2

B. Support Vector Machine Classification

The support vector machines are mainly proposed for supervised learning such as classification and regression [4, 7]. SVM classification supports both binary and multiclass targets. To apply a SVM classifier, there are two important steps: one is feature selection (new features are selected from the original inputs), another is the reduction of the training dataset. The goal of the SVM approach is to define a hyperplane in a high-dimensional feature space Z, which divides the set of samples in the feature space such that all the points with the same label are on the same side of the hyperplane (Fig.2). The surface is often called the optimal hyperplane, and the data points closest to the hyperplane are called support vectors.

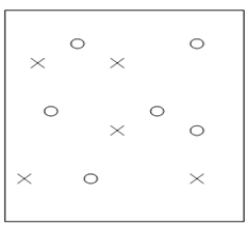

Fig.2 (a). Training data

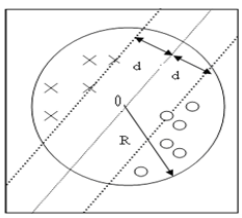

Fig.2 (b). Hyperplane

Fig.2(a) represents training set in the input space, symbols "x" and "o" represents different classes. In Fig.2 (b) represents transformed training set in a high-dimensional feature space.

SVM Kernel Function:

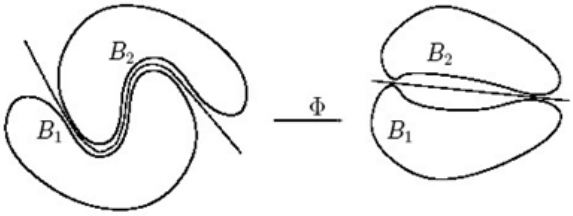

Figure 3. Kernel Function – Hyperplane

The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation (Fig.3). It is used to increase the computational power of the linear learning machines [8].

Polynomial Kernel:

The kernel function is defined as,

$$K(x_i, x_j) = (1 + x_i^T x_j)^p \qquad \text{----- (1)}$$

where, *p* is the degree of the polynomial. The motivation is that in general, for vectors xi that are linearly dependent on *p* dimensions, the kernel function of order *p* can be used to transform them into linearly independent vectors on those *p* dimensions (Eq.1). Once they are transformed into the dimension space where they become linearly separable, the linear- SVM case can handle the classification problem.

Radial Basis Function (RBF) Kernel:

The kernel function is defined as,

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \qquad \text{------ (2)}$$

This kernel is basically suited best to deal with data that have a class-conditional probability distribution function approaching the Gaussian distribution [10] (Eq.2). It maps such data into a different space where the data becomes linearly separable.

C. Feature Selection Using Genetic Algorithm

Integer – Coded Genetic Algorithm:

Integer-coded genetic algorithms (ICGA) devised by Holland [11] have been applied to many optimization problems with success. ICGA select top N best features out of total M features for classification. It is used to select important and relevant features. Discard the irrelevant and redundant features. ICGA proceeds with initial population as 50 chromosomes. Each chromosome consists of array size as '*N*' and can be arranged randomly from 1 to *N* for the corresponding feature subset.

The initial population is generated to *Tournament Selection*

where the fitness values of consecutive pairs of chromosomes are compared and the winner is selected for Crossover operation. Crossover operation is followed by the mutation operation which maintains the diversity from one generation of population to next by randomly changing a gene sequence of a chromosome. Then the mutation site is generated for each chromosome and the value will be replaced randomly for the generated integer from 1 to N. From randomly generated chromosomes the repetitions will be removed as shown in Figure 4.

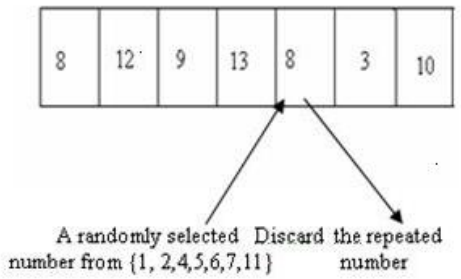

Figure 4. Remove Repetition.

This operation is performed for each chromosome and replaced by a randomly selected number from the set (Z_n – G_i), where Z_n is the set of integers from 1 to *N* and G_i is the set of numbers present in the i^{th} chromosome.

Binary – Coded Genetic Algorithm:

Binary-coded genetic algorithms (BGA) devised by Holland [11] have been applied to many optimization problems with success. When an optimization problem involves binary strings of extra length due to a large number of design variables or unusually high precision on the design variable representations, the number of representative grids in the design space becomes very large, often resulting in search difficulties.

There are many types of crossover strategies in BGA. In two pairing parental binary strings, two bit locations are randomly selected, and the substrings between these two bit locations on the parental strings exchange with each other, thereby forming two new offspring strings. Two parental binary strings and their offspring binary strings after the crossover operation are shown in (Figure 5 and 6) in which the 5th bit and the 11th bit, both underlined, are two randomly picked sites for crossover.

1011**0111001**0100

1000**1010111**0010

Figure 5. Parental binary strings (before crossover)

1011**1010111**0100

1000**0111001**0010

Figure 6. Offspring binary strings (after crossover)

Steps in GA:

The Figure 7. shows the different steps used in GA.

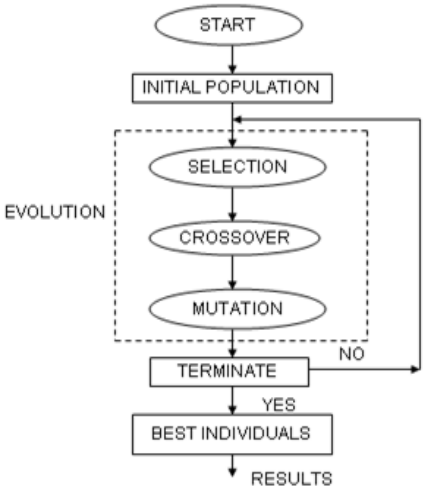

Figure 7. Steps in GA

III. EXPERIMENTAL RESULTS

A. Data Preprocessing

In real world database are highly susceptible to noisy, missing and inconsistent data due to huge size (often several gigabytes or more) data. In Preprocessing data's are organized such that each column contains expression levels of different values of Wisconsin Breast Cancer Database. Data cleaning (or Data Cleansing) is a technique used in preprocessing. It attempts to fill in the missing values, smooth out noise while identifying outliers and correct inconsistencies in the database. In Wisconsin Breast Cancer Dataset there are 16 missing attributes. Each of the missing values are analyzed with respect to its corresponding class label and is replaced by the most frequent occurrence in its class. The results of missing value are tabulated in III.

Table III. Data Preprocessing Results

S.No.	Attribute Name	No. of Missing Values		Maximum Occurrence		Replaced Value	
		2	4	2	4	2	4
1	Bare Nuclei	14	2	1(387)	10(129)	1(401)	10(131)

B. Support Vector Machine

Accuracy and Execution Time:

The dataset is fed into the support vectors, accuracy value and execution time is recorded for various kernel functions such as Polynomial and RBF. The kernel function, accuracy value and execution time is recorded and shown in tabulated IV.

Table IV. Results of Accuracy and Execution Time

Kernel Function	Accuracy (%)	Time taken (ms)
RBF	83.1325	4031
Polynomial	83.1325	12700

C.Feature Selection Using Genetic Algorithm ICGA and BGA

In this module we have analyzed the classification accuracy and execution time for ICGA and BGA with SVM Polynomial and RBF kernel function. The result is recorded and shown in Table V.

Table V. Results of Accuracy and Execution Time for Genetic Algorithm

Genetic Algorithm	Kernel Function	Accuracy (%)	Time taken (ms)
ICGA	RBF	84	36
	Polynomial	83	36
BGA	RBF	85	31
	Polynomial	87	31

D.Comparison of ICGA and BGA

Figure 8. shows the compared graph for Integer and Binary Coded Genetic Algorithm with SVM Polynomial and RBF. The graph shows that Binary Coded Genetic Algorithm gives better classification accuracy than the Integer Coded Genetic Algorithm for both SVM RBF and Polynomial functions.

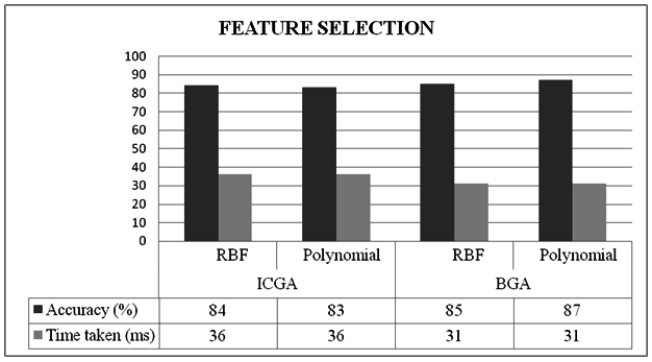

Figure 8. Comparison Chart for Classification Accuracy and Execution Time

IV. CONCLUSION AND FUTURE WORK

The goal of this thesis is to design Support Vector Machine, Integer and Binary Coded Genetic Algorithm were analyzed to find the classification accuracy and runtime for various kernel functions such as Polynomial and Radical Basic function are used. Feature Selection algorithm is used to improve the classification accuracy of classifier with respect to medical datasets. In this paper, a new algorithm Binary Coded Genetic Algorithm is implemented to induce classification accuracy from the training dataset. Compared to previous implementation, Binary Coded Genetic Algorithm gives better than Integer Coded Genetic Algorithm and the execution time will be same for both the Genetic Algorithms using various kernel functions. The future work can be implemented using any other kernel function in Support Vector Machine and can also extend for a larger database to get more accurate results for classification. It can also extend to time variant data sets and also for Regression method.

REFERENCES

[1] Sumit Bhatia, Praveen Prakash, and G.N. Pillai, "SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features", WCECS 2008, October 22 - 24, 2008.
[2] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2007.
[3] Cheng-Lung Huang , Chieh-Jen Wang," A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications 31 (2006) 231–240.
[4] Vapnik, V. N. The nature of statistical learning theory. New York: Springer, 1995.
[5] Li Zhuo, Jing Zheng, Fang Wang, Xia Li, Bin Ai, Junping Qian, "A Genetic Algorithm Based Wrapper Feature Selection Method for Classification of Hyperspectral Images Using Support Vector Machine**",** The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing 2008.
[6] UCI Machine Learning Repository: Breast Cancer Data Set. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Original.
[7] C.J.C.Burges, "A tutorial on support vector machines for pattern recognition.Data mining and knowledge discovery", 2(2): 121-167, June 1998.
[8] Data Mining and Knowledge Discovery", 2(2): 121-167, June 1998.
[9] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf, "An Introduction to Kernel-Based Learning Algorithms", IEEE Transactions on Neural Networks, VOL. 12, NO. 2, MARCH 2001.
[10] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf, "An Introduction to Kernel-Based Learning Algorithms", IEEE Transactions on Neural Networks, VOL. 12, NO. 2, MARCH 2001.
[11] Holland, J. H. (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, Michigan
[12] Ying Tan and Jun Wang, "A Support Vector Machine with a Hybrid Kernel and Minimal Vapnik-Chervonenkis Dimension", IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 4, April 2004.
[13] Stefan Lessmann, Robert Stahlbock, Sven F. Crone, "Genetic Algorithms for Support Vector Machine Model Selection", International Joint Conference on Neural Networks, July 16-21, 2006

BIBLIOGRAPHY

D.Nithya graduated from Avinashilingam University, in Computer Science and Engineering during the year 2008. She obtained her Master degree in Computer Science and Engineering from Avinashilingam University, Coimbatore in the year 2010. At present she is an assistant professor in the Department of Computer Science and Engineering, Faculty of Engineering, Avinashlingam University for Women, Coimbatore, India. Her area of interest includes Data mining and Neural networks. She has 3 years of experience in teaching**.**

V.Suganya graduated from Bharathiyar University, in Electrical and Electronics Engineering during the year 2002. She obtained her Master degree in Computer Science and Engineering from Anna University of Technology, Coimbatore in the year 2010. At present she is an assistant professor in the Department of Computer Science and Engineering, Faculty of Engineering, Avinashlingam University for Women, Coimbatore, India. Her area of interest includes Data mining and Wireless Networks. She has 3 years of experience in teaching.

R.Saranya Irudaya Mary graduated from Avinashilingam University, in Information Technology during the year 2008. She obtained her Master degree in Computer Science and Engineering from Avinashilingam University, Coimbatore in the year 2010. At present she is an assistant professor in the Department of Computer Science and Engineering, Faculty of Engineering, Avinashlingam University for Women, Coimbatore, India. Her area of interest includes Data mining and Image Processing. She has 3 years of experience in teaching.