Optimize Support Vector Machine Classifier based on Evolutionary Algorithm for Breast Cancer Diagnosis

Riyadh AbdEl-Salam Hassan

Computer science department, College of computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Egypt.

AbdEl-Fatah Hegazy

College of computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Egypt.

Amr Ahmed Badr

Computer Science department, Faculty of Computers and informatics, Cairo University, Egypt.

Summary
Support Vector Machine has become the reference for many classification problems, because of their flexibility, and capacity to handle high dimensional data. However, the success of Support Vector Machine classifier depends on the perfect choice of the values of its parameters along with the feature subset selection. So the objective of this paper is to propose an Evolutionary Optimization Algorithm (EA) for feature selection and parameter optimization to solve this kind of the SVM depended to improve its performance accuracy. The proposed approach is compared with other approach. The results show that our approach obtained the highest classification accuracy (100%) with limit feature subset selected.

Key words:
Support Vector Machine; Evolutionary Algorithm; Breast Cancer; Feature Selection; Parameter Optimization.

1. Introduction

Support Vector machine (SVM) were first suggested by V. Vapink [1] and are the most powerful of the Classification, Regression, Bioinformatics, Pattern Recognition, etc. Support Vector Machine has the ability to obtain a global optimal solution by the ability to separate the samples from different classes by a hyperplane with largest margin. Therefore, two problems are encountered when using SVM: how to select optimal feature subset inputs for SVM and how to set the best kernel parameters.

These two problems are crucial to taking into account because the feature subset choice influences the appropriate kernel parameters, therefore in the accuracy of diagnosis [2]. So, obtaining the optimal feature subset and SVM parameters must occur simultaneously.

Feature selection is an important issue in building SVM classifier, and data mining community. The purpose of feature selection subset is to limit the number of input features in order to have a good classification performance [3].
In this work, we propose a SVM for diagnosis the breast cancer, and Genetic Algorithm for optimize all properties of the SVM classifier. WBCD data set was taken from the University of California at Irvine (UCI) for training and testing experiments.

This paper is organized as follows: Introduction to the SVM is given in section 2. Section 3 describes the basic EA concepts. Section 4 describes the EA based feature selection and parameter optimization. Section 5 presents the experimental result from using the proposed method. Section 6 gives out the conclusions2. Tables, Figures and Equations

2. Introduction to Support Vector Machines

Suppose a training data set with input data vector $x_i \in R^n$ and corresponding class labels $y_i \in \{-1, +1\}, i = 1,2, \dots, m$ is given, the idea of Support Vector Classification is to compute a hyperplane that maximize the margin separating between the two class of samples [4]. This hyperplane is shown in Fig.1.
A separating hyperplane is determined by

$$(w.x_i) + b \geq +1 \quad for \quad y_i = +1 \qquad (1)$$
$$(w.x_i) + b \leq -1 \quad for \quad y_i = -1 \qquad (2)$$

Where $(w.x_i)$ shows the inner product of w and x_i, the inequalities in Eqs. (1) and (2) can be combined as in

$$y_i((w.x_i) + b) - 1 \geq 0 \qquad \forall_i = 1,2,3,\dots\dots,m. \qquad (3)$$

The SVM classifier finds an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

Manuscript received December 5, 2011
Manuscript revised December 20, 2011

$$\min_{w,b} \frac{1}{2} w^T.w \quad subject\ to\ y_i((w.x_i)+b) \geq 1 \qquad (4)$$

This optimization problem is named as quadratic optimization problem. For solving of this problem, we must find the saddle point of the Lagrange function:

$$L_p(w,b,\alpha) = \frac{1}{2} w^T.w - \sum_{i=1}^{m} (\alpha_i y_i((w.x_i)+b) - 1) \qquad (5)$$

$$\text{subject to } \alpha_i \geq 0$$

Where *w* represents a vector that defines the boundary, x_i is input data points, *b* represents a scalar threshold value, and α_i is Lagrange multiplier and it must be ≥ 0. L_p function should be minimized with respect to the *w* and *b* variables, and maximized with respect to the non-negative dual variable α_i [5]. When the L_p function performs derivative, the following equations are obtained:

#

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \quad for \quad \frac{\partial}{\partial w} L_p = 0 \qquad (6)$$

$$\sum_{i=1}^{m} \alpha_i y_i \quad for \quad \frac{\partial}{\partial b} L_p = 0 \qquad (7)$$

The Karush Kuhn-Tuker (KKT) situations for the optimum constrained function are necessary for a maximum of Eq. (5) [6] [7].

If Eqs. (6) and (7) are substitute into Eq. (5), then L_p is changed to the dual Lagrangian $L_D(\alpha)$:

$$\max_{\alpha} L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i.x_j) \qquad (8)$$

$$subject\ to\ \forall_i: \alpha_i \geq 0, and \sum_{i=1}^{m} \alpha_i y_i = 0$$

The α_i Lagrange multiplier is calculated by using:

$$\alpha_i(y_i((w.x_i)+b) - 1) = 0 \qquad (9)$$

The dual Lagrangian $L_D(\alpha)$ must be maximized with respect to non-negative α_i. This is a standard quadratic optimization problem. For solving this dual optimization problem by determines the appropriate parameters *w* and *b* of the optimal hyperplane with respect to α_i.the optimal decision hyperplane $f(x)$ can be given as:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i (x_i.x) + b \qquad (10)$$

The respective training vectors are called support vector, when input data point x_i has a non-zero Lagrange multiplier α_i.

The previous concept can be extended to the non-separable case by replace Eqs. (1) and (2) as:

$$(w.x_i) + b \geq +1 - \xi_i \quad for \quad y_i = +1 \qquad (11)$$

$$(w.x_i) + b \leq -1 + \xi_i \quad for \quad y_i = -1 \qquad (12)$$

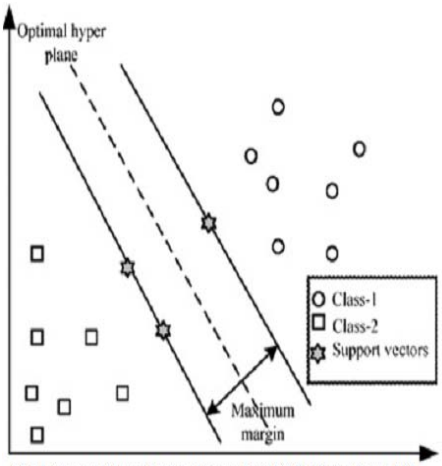

Fig.1. Separation of two classes by SVM, Photo from [8].

Where ξ_i is a non-negative slack variable and $\xi_i \geq 0, i = 1,2, \ldots \ldots, m$. this slack variable keeps the constraint violation as small as and provides the minimum training error [5]. Thus, Eq. (4) can be changed as

$$\min_{w,b,\xi} \frac{1}{2} w^T.w + C \sum_{i=1}^{m} \xi_i \qquad (13)$$

$$subject\ to \quad y_i((w.x_i)+b) + \xi_i - 1 \geq 0, \xi_i \geq 0$$

So this optimization model can be solved as separable case by using the Lagrangian method

$$\max_{\alpha} L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i.x_j) \qquad (14)$$

$$subject\ to\ \forall_i: 0 \leq \alpha_i \leq C, i - 1,2,\ldots,m \ \ and \sum_{i=1}^{m} \alpha_i y_i - 0$$

Where C is soft margin constant penalty parameter, which is determined by the user [5][6]. The optimal decision hyperplane function $f(x)$ is the same as Eq. (10).

When a linear boundary is not used, SVM can map input vector training samples in input space to higher-dimensional feature space via the kernel function φ. Thus, the inner products in the optimizing dual Lagrangian $L_D(\alpha)$ Eq. (8) are substituted by the kernel function as

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i . x_j) \quad (15)$$

$$subject\ to\ \forall_i: 0 \le \alpha_i \le C, i = 1,2,\ldots,m \ \ and \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$where \ \left(\varphi(x_i) . \varphi(x_j)\right) = k(x_i . x_j)$$

So the optimal decision hyperplane $f(x)$ of a non-linear SVM can be shown as

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i k(x_i, x) + b = \sum_{i \in sv} y_i \alpha_i k(x_i, x) \quad (16)$$

Where *SV* is support vector number.

3. The basic Concepts of Evolutionary Algorithm

EAs are optimization algorithms that use biology-inspired techniques, such as mutation, crossover, natural selection, and survival of the fittest [9]; and work with a set of Candidate solution (Populations) which are represented by chromosomes. Each chromosome comprises of a sequence of individuals structures called genes which represent the actual parameters to be optimized. Fig.2 shows the principle of all EAs proceeds.

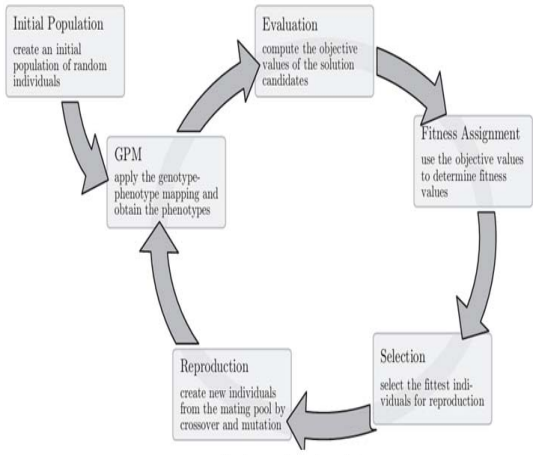

Fig.2. Basic Cycle of EAs, photo from [9].

The Genetic Algorithm (GA) is a type of evolutionary algorithms where the solutions are represented in binary strings of 0s and 1s or arrays of other types. The GA begin generate a new solutions through a set of operators. The first operator is reproduction; it reflects the principle of survival of the fittest, the survival solutions are copied to the next generation based on the quality of their objective function value (Fitness function). Crossover is the second operator, which is a random technique for exchanging genes between two chromosomes using the one point crossover, two point crossovers, or homologue crossover. The third operator is mutation where the genes may occasionally be altered [5] [10]. The new populations after done the three operators of GA called offspring, and the main objective of GA that the offspring will be better than the old population. Fig.3. show the genetic operator of crossover and mutation.

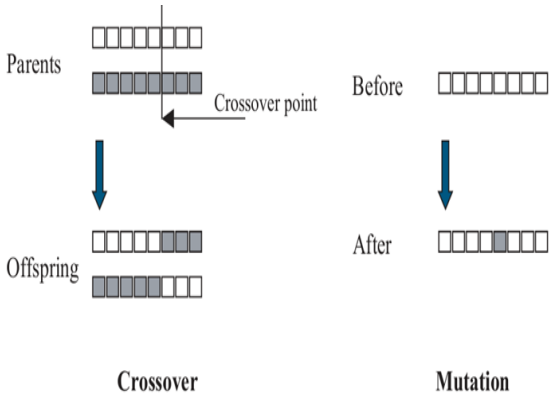

Fig.3. Crossover and Mutation Operators.

4. EAs based Feature Selection and Parameter Optimization

4.1. Chromosome Design

This proposed approach will be used two different Kernel functions RBF (Radial Basis Function) and sigmoid kernels. These kernel functions have parameters that always significantly influence the performance. For instance, the parameter C of SVM has an important influence on its performance. So we need to be carefully selected its value.

When one of these kernels is selected, the parameters of it, split- ratio validation, and features used as input attributes must be optimized using our approach. Fig.4. shows the design chromosome of these kernels.

The chromosome is divided into three parts as previous shown. The first part represents the parameters of the kernel function that will be optimized, # $g_1^c \ldots g_{m_c}^c$ represents binary coding of parameter C, $g_1^\gamma \ldots g_{m_\gamma}^\gamma$ is binary coding of γ, $g_1^{c0} \ldots g_{m_{c0}}^{c0}$ is binary coding of coef0, and m_c, m_γ, m_{c0} represents the number of bits in binary code of **C**, γ, and coef0 parameters. The second part represents binary coding of split ratio validation to determine the optimal splitting ratio of dataset from 50% to 80% for training data. The last part

represents binary coding of the feature subset selection, in this coding of feature subset, '1' represents that feature is selected; '0' represents the feature is not selected.

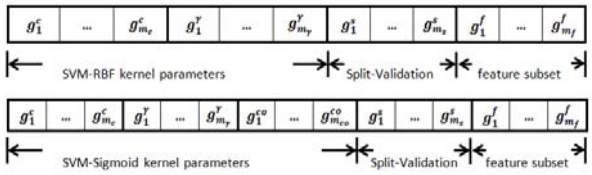

Fig.4. SVM-Kernels parameters, Split Validation and Feature Selection Chromosome

4.2. Feature Selection

The main advantage of feature selection strategy is to limit the number of input feature in classifier in order to have high performance accuracy.

Feature selection methods fall into two types: the filter Methods and the wrapper methods [12]. The Filter methods select subset of features as a pre-processing step, independently of the chosen classifier, but the Wrapper methods selected the features according to their classifier performance. GA-SVM approach is applying the concepts of the two methods. Fig.5. shows the meaning of them.

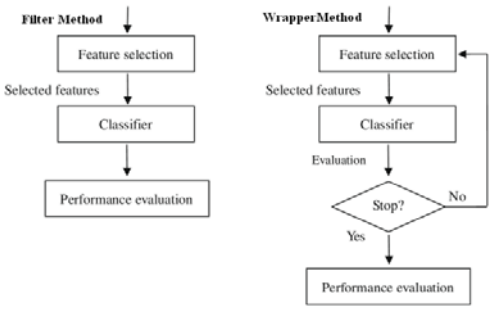

Fig.5. Filter and Wrapper Feature Selection Methods

4.3. Fitness Function

Fitness function assessing the performance for each chromosome and must be designed before selects the optimal values of SVM classifier parameters, and feature subset. There are several indicators employed to evaluate the prediction accuracy of models such as accuracy rate, error rate, precision rate, and recall rate. In this paper accuracy rate, sensitivity, specificity, positive predictive value, and negative predictive value will be used as majors to evaluate the chromosome fittest.

Table (1) contains information about actual and predicted classification done by classifier for two classes. The classification accuracy of each chromosome are computed by using TP+TN/N. where TP (TP = true positive fraction) is correctly classified as positive, TN (TN = true negative fraction) is correctly classified as negative, FP (FP = false positive fraction) is a case with negative class will be classified as positive, and FN (FN = false negative fraction) is a case with positive class will be classified as negative. N (TP+TN+FP+FN) is total number of testing dataset samples. Sensitivity (TP/TP+FN) is the rate of cases with positive class that are classified as positive. Specificity (TN/TN+FP) is the rate of cases with negative class that are classified as negative. Positive predictive value (TP/TP+FP) is the rate of cases with positive class that correctly classified. Negative predictive value (TN/TN+FN) is the rate of cases with negative class that correctly classified.

Table (1): Confusion Matrix Representation.

Actual Value	Prediction outcome	
	Class B	Class M
Class B	TP	FN
Class M	FP	TN

4.4. GA-SVM Approach

Fig.6. show the main steps of the proposed GA-SVM approach. The Explanation about the steps is as follows:

1- Generating. This step will transfer each parameter and feature subset from its genotype genetic code into a phenotype.
2- Split-Validation. After converting the split-ratio parameter, the Split Validation will divide the dataset into training and testing dataset according to split-ratio value. This split-ratio will take random value from 50% to 80% for the training data, and the rest for testing.
3- Feature Subset. After step 1, feature subset can be determined which feature will be selected.
4- Training SVM Classifier. SVM classifier is trained by training set with feature subset selected and variable value of parameters.
5- Evaluation (Fitness). For each chromosome representing, the testing dataset is used to calculate the classification accuracy of classifier. Then each chromosome is evaluated according to the indicators of the fitness function.
6- Satisfy terminal. If the population does not satisfy the termination condition, we proceed with the next generation operation. The termination criteria are that the max generation number reached or the fitness function value does not improve during the last 15 generations.
7- Genetic Operation. The reproduction operators selected 20% of the best chromosome

performance to reproduce the new offspring chromosome, and then applying crossover and mutation operators.

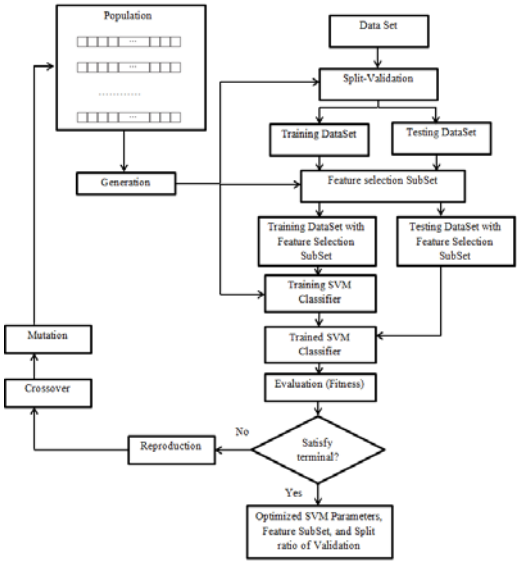

Fig.6. GA-SVM Approach

5. Experiment results

5.1. Breast Cancer DataSet (WDBC)

We have used the commonly dataset used among researchers who apply machine learning methods for breast cancer classification [11]. This dataset contains 683 samples taken from needle aspirates from human breast cancer tissue, that distributed to 444 samples belong to benign class, and the rest are of malignant class. Also it consists of 10 attributes, each of one represented as an integer value between 1 and 10, and the last attribute are label nominal value B for Benign and M for Malignant. The attributes are; clump thickness (F1), uniformity of cell size (F2), uniformity of cell shape (F3), marginal adhesion (F4), single epithelial cell size (F5), bare nucleoli (F6), bland chromatin (F7), normal nuclei (F8), and mitoses (F9).

5.2. Experiment descriptions

In this section, we evaluate the proposed GA-SVM approach to predict Breast Cancer Diagnosis and compared with others approach. We decide to select two types of kernels RBF and Sigmoid for implementing SVM and show its results.

The details parameter setting for genetic algorithm is as the following: tournament selection type, crossover rate 0.5, mutation rate -1.0, and uniform crossover type. For kernel parameters it will be generated randomly 0:1, except coef0 parameter from -1:1.

Table (2) shows the optimal parameter values for GA-SVM approach and classification accuracy on the testing dataset. Among the model running it achieved the highest classification accuracy when selected four feature subsets for the 76.7-23.3% training-testing partition for RBF kernel, and selected 7 feature subset for the 80-20% training-testing for sigmoid kernel.

Also we present values of sensitivity, specificity, positive predictive value and negative predictive value for the two kernels in Table (3).

In this study, there are two classes as Benign (B) and Malignant (M).Table (4) and Table (5) shows the classification results of the RBF and sigmoid kernels using confusion matrix.

For comparison, Table (6) gives the classification accuracies of our approach with two kernels, and previous approaches. As we can see from the results, our GA-SVM Approach obtains the highest classification accuracy.

Table (2): Optimal parameters and Feature subset selection

Approach	C	γ	Coef0	Split-Ratio (%)	No. Feature subset	Feature-Name	Accuracy (%)
RBF	1	0.329	-	76.7	4	F1,F6,F7,F8	100
Sigmoid	0.283	0.003	0.728	80	7	F1,F2,F3,F4,F6,F7,F9	100

Table (3): Indicator values for RBF and Sigmoid Kernel

Measures	GA-SVM RBF	GA-SVM Sigmoid
Sensitivity (%)	100	100
Specificity (%)	100	100
Positive predictive value (%)	100	100
Negative predictive value (%)	100	100

Table (4): Confusion Matrix for RBF testing

Actual Value	Prediction outcome	
	Class B	**Class M**
Class B	97	0
Class M	0	62

Table (5): Confusion Matrix for sigmoid testing

Actual Value	Prediction outcome	
	Class B	**Class M**
Class B	82	0
Class M	0	55

Table (6): Classification accuracies obtained from different approach

Approach	Year	Classification Accuracy (%)
F-Score+SVM [13]	2007	99.51
LS-SVM [14]	2007	98.53
GA-SVM RBF	This study	100
GA-SVM Sigmoid	This study	100

6. Conclusion

This paper propose Evolutionary optimization algorithm, which can optimize the parameter values for SVM, and obtain the optimal subset of features. A comparison of the obtained results with other approach demonstrates that the

proposed approach improves the classification accuracy rates. The GA-SVM approach was applied to remove insignificant features and effectively find best parameter values.

The results show that the classification accuracy of GA-SVM is the highest of other approach. Therefore, the approach can be used for many classification areas or can be used for Breast Cancer Prognosis using these two kernels or other kernels.

References

[1] Vapnik, V. (2000). The Nature of Statistical Learning Theory. New York: Wiley.

[2] Frohlich, H., Chapelle, O., Scholkopf, B. (2003). Feature Selection for Support Vector Machines by Means of Genetic Algorithms. 15th IEEE International Conference, 142-148.

[3] Zhang, G.P. (2000). Neural Network for a Classification: a survey. IEEE Transactions on Systems, Man and cybernetics-part C: Applications and Reviews, 30(4), 451-462.

[4] Cristianini, N., Shawe-Taylor, J. (2000). An introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

[5] Hung, C., Wang, C. (2006). A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications, 31, 231-240.

[6] Fernandez Pierna, J. A., Baeten, V., Michotte Renier, A., Cogdill, R. P., Dardenne, P. (2004). Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. Journal of Chemometrics, 18, 341-349.

[7] Gunn, S. R. (1998). Support vector machines for classification and regression. Technical report, University of Southampton, UK.

[8] Aydin, I., Karakose, M., Akin, E. (2011). A multi-objective artificial immune algorithm for parameter optimization in support vector machine. Applied Soft Computing, 11, 120-129.

[9] Weise, T. (2009). Global Optimization Algorithms -Theory and Application-. Available at: http://www.it-weise.de/ .

[10] Hajri-Gabouj, S. (2002). A Fuzzy Genetic Optimization Algorithm for an assignment problem. Systems, Man and Cybernetics, 2002 IEEE International Conference. 3, 6pp.

[11] Wolberg, W.H. (1992). Wisconsin Breast Cancer Database, University of Wisconsin Hospitals Madison, Wisconsin, USA. Available-at: http://mlr.cs.umass.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[12] Guyon, I., Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

[13] Akay, M. (2009) Support Vector Machines combined with feature selection for breast cancer diagnosis. Expert Systems with Application, 36, 3240-3247.

[14] Polat, K., Gunes, S. (2006). Breast Cancer Diagnosis using least square Support Vector Machine. Digital Signal Processing, 17, 694-701.

Riyadh AbdEl-Salam received the BSc of computer science degree from Suez Canal University, Egypt.

AbdEl-Fatah Hegazy, Prof.Dr, is working as dean assistant Post-graduate studies at Arab Academy for Science, Technology and Maritime Transport, Cairo branch.

Amr Badr, Prof.Dr, at Computer Science department, Faculty of Computers and informatics, Cairo University, Egypt.