# Feature Selection Using Multi Objective Genetic Algorithm with Support Vector Machine

Mojgan Elikaei Ahari [a], Babak Nasersharif [b,*]

[a] *Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
[b] *Electrical and Computer Engineering Department, K.N. Toosi University of Technology, Iran*

Received 1 December 2015; accepted 29 January 2016

**Abstract**

Different approaches have been proposed for feature selection to obtain suitable features subset among all features. These methods search feature space for feature subsets which satisfies some criteria or optimizes several objective functions. The objective functions are divided into two main groups: filter and wrapper methods. In filter methods, features subsets are selected due to some measures like inter-class distance, features statistical independence or information theoretic measures. Even though, wrapper methods use a classifier to evaluate features subsets by their predictive accuracy (on test data) by statistical resampling or cross-validation. Filter methods usually use only one measure for feature selection that does not necessarily produce the best result. In this paper, we proposed to use the classification error measures besides to filter measures where our classifier is support vector machine (SVM). To this end, we use multi objective genetic algorithm. In this way, one of our feature selection measure is SVM classification error. Another measure is selected between mutual information and Laplacian criteria which indicates informative content and structure preserving property of features, respectively. The evaluation results on the UCI datasets show the efficiency of this method.

*Keywords:* feature selection, multi objective genetic algorithm, support vector machine.

## 1. Introduction

Classification is to assign a data to a (specific) category. At first we need a classification system based on the input data that can determine their categories. Another case is about the information extracted from data called features which classifiers use them for assigning data to categories. So, the patterns existed in data are represented by a feature vector.

For better classification results, we should use useful feature vectors with suitable dimension which can discriminate well between data classes. To this end, many feature selection and transformation and dimension reduction methods have been proposed [1]. There are two groups of method in feature selection: filter and wrapper methods. In filter methods, features subsets are selected due to some measures like inter-class distance, features statistical properties or information theoretic measures. Even though, wrapper methods use a classifier to evaluate features subsets

* Corresponding author. Email: bnasersharif@eetd.kntu.ac.ir

with their predictive accuracy (on test data) by statistical resampling or cross-validation. [2]

It also filters and wrapper methods have been developed recently and feature selection is considered as multi-objective problem.[3]

Conventional feature selection techniques usually demand many samples to estimate statistics accurately. In addition, they are usually based on an exhaustive process for finding the best set of features, and in this case, they are time demanding, and their CPU processing time exponentially increases as the number of bands (features) increases. To this extent, a new generation of feature selection techniques is based on evolutionary optimization methods, since they are not based on an exhaustive process and can lead to a conclusion in a faster way. In addition, by considering an efficient fitness function for these methods, they can handle high-dimensional data with even a limited number of training samples. [4]

Support vector machine (SVM) is a supervised method, which is used for classification and regression. Different approach was proposed for feature selection for support vector machine to enhance the quality of feature selection by reducing the search space and time calculations. Some feature selection methods are without relying on SVM it means they selected the important features and then SVM is used for classification. Studies have shown that support vector machine can not directly extract the importance of a feature. [3]

Previous proposed feature selection methods usually select or transform features without attention to SVM classification and training criteria which may tends to accuracy degradation of SVM. In the newer methods feature performed in a way that classification accuracy has been preserved. One of the most effective methods of feature selection for SVM is the method that feature selection Turned to a model of SVM, unlike hybrid search methods [4]. In some other feature selection methods, evolutionary algorithms with SVM accuracy as their fitness function have been used and the results show that these methods lead to increasing, classification accuracy and reducing computational complexity [5]. Sometimes fitness function is considered as a linear combination of recognition accuracy and the member of selected features. In some methods, Meta heuristic algorithms and the border on SVM generalization error have used for feature selection criteria. [6]

In this paper, we propose to use the classification error measures along with filter measures where our classifier is SVM. To this end, we use multi-objective genetic algorithm where one of our objective for feature selection is SVM classification error. On the other hand, another objective is selected among mutual information and Laplacian criteria which indicates informative content and structure preserving property of features, respectively.

The remainder of this work is organized as follows. Section 2 discuses feature selection criteria brifely Section 3 includes proposed method based on multi objective genetic algorithm. Section 4 reports evaluation results of proposed method. Finally, our conclusions have been given in Section 5.

## 2. The Criteria and Their Usage in Feature Selection

The filter methods, don't attend to the classification method and training algorithm. They don't depend on applied machine learning algorithm and evaluate features subset using other criteria.

Three steps for Filter methods are as follows: [9]

1- A rank is calculated for each feature (by a threshold)

2- This rank is then arranged and features with the lowest rank will be removed

3- As an input, selected high rank features are given to classification system

In the following sub-sections, we describe two used filter methods in this paper.

### 2.1. Mutual Information

Mutual information is defined as information shared between two random variables. How much information can be given about random variable Y by random variable X.

Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X,Y) = -\sum_{x\in X, y\in Y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

Where p(x,y) is the joint probability distribution function. Mutual information *I(X,Y)*, has a large value if two variables X and Y are closely related to each other. Otherwise, if X and Y will be completely independent, *I(X,Y)*, will be zero. In filter methods, Mutual information is used to measure the relationship between selected features and class labels. [10]

The mutual information (MI) is a measure of the amount of information that one random variable has about another variable. This definition is useful within the context of feature selection because it gives a way to quantify the relevance of a feature subset with respect to the output vector C. [11]

The main purpose of using mutual information criterion is maximizing information between classes and features and removing less informal features.

### 2.2. Laplacian Measure

The Laplacian criterion is used for finding a transformation from a high dementional space to alower dimentional space such that preserve the minimum local properties. So it is based on minimizing the distance between data samples in lower dimentional spaces considering their distance in higher dimentional space. Mapping input data X, in a d-dimensional space is the goal of this method. This mapping is such that locally communication is preserved in nearby neighbourhoods. [8]

The calculation steps are as follows:

1: Creating the adjacency graph by k-means

2: Weighting graph edges using heat kernel

$$W_{i,j} = e^{-\frac{\|x_i - x_j\|^2}{t}} \qquad (2)$$

3: (Eigenmaps) Compute eigenvalues and eigenvectors for the generalized eigenvector problem:

$$Lf=\lambda Df \qquad (3)$$

Where D is diagonal weight matrix, its entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_j W_{ij}$, L=D-W is the Laplacian matrix. Laplacian is a symmetric, positive semidefinite matrix which can be though of as an operator on functions defined on vertices of graph G.

Let $f_0,\ldots,f_{k-1}$ be the solutions of equation 3, ordered according to their eigenvalues,

$Lf_0=\lambda_0 Df_0$
$Lf_1=\lambda_1 Df_1$
…
$Lf_{k-1}=\lambda_{k-1} Df_{k-1}$
$0=\lambda_0 <= \lambda_1 <= \ldots. <= \lambda_{k-1}$

We leave out the eigenvector $f_0$ corresponding to eigenvalue 0 and use the next m eigenvectors for embedding in m-dimentional Euclidean space [9].

$$X_I \longrightarrow (f_1(i),\ldots, f_m(i))$$

## 3. Our Method

As mentioned above, in filter methods, features subsets are selected due to some measures like inter-class distance, features statistical properties or information theoretic measures that only depends on data. Therefore, filter methods generally consider a data dependent criterion without attention to the used classification, method. In this paper, we propose to consider filter criteria methods, alone with classification error criteria for feature selection. To

this end, we use multi-objective genetic algorithm where SVM is our classifier and its error is one of our objective. As another objective, we use mutual information and Laplacian criteria which indicate informative content and structure preserving property of features, respectively.

### *3.1. Initialization*

Our defined chromosome is a binary string of length d(dimension of feature vector) where 1 in string denote the selection of feature and 0 indicate the removed features. We generate a random initial population where number of initial population is npop. (here npop=100)

### *3.2. Determine Fitness of Population*

The fitness function evaluates the quality of solutions. We used a fitness function which considers both between mutual information or Laplacian and SVM classification error for obtaining a higher classification rate. For this purpose, at first, we define two different functions.

The first fitness function is based on criteria of mutual information or Laplacian. We used objective functions introduced in relations (1) or the Laplacian matrix, as the first evaluation function.

The second evaluation function is based on SVM classification error rate which should be minimized. It is calculated as ratio of correctly classified test samples number to the total number of test samples.

$$Classification\_accuracy = \frac{N_{correct}}{N} \qquad (4)$$

Where N is the total number of test samples and, $N_{correct}$ is the number of samples detected.

### *3.3. Non Dominated Sor Method*

Once the population is initialized, it is sorted based on non-domination into each front. The first front is completely non-dominant set in the current population and the second front is being dominated by the individuals in the first front only and so on. Each individual in the each front have been assigned rank values based on front in which they belong to. Individuals in first front are given a fitness value of equal to 1 and individuals in the second are assigned fitness value as equal to 2 and so on. [10]

### *3.4. Crowding Distance*

Once the non-dominated sort is perfect, the crowding distance is determined. We selected the individuals based on rank and crowding distance, all the individuals in the population have been assigned a crowding distance value. Crowding distance is determined in a front-wise method. Thus, comparing the crowding distance between two individuals in different fronts is meaning-less

The crowding distance between two offspring can be defined as:

$$d_j(k) = \sum_{i=1}^{n} \frac{f_i(K-1) - f_i(K+1)}{f_i^{\max} - f_i^{\min}} \qquad (5)$$

Where $f_i(k-1)$ is the i-th objective function's crowding distance with k-th offspring. [10]

### *3.5. Selection Operator*

This operator chooses from population chromosomes, the number of chromosomes to generate the next generation. The fittest chromosomes have a higher chance to be selected for next generation. Here two parents are chosen based on the roulette wheel.

### *3.6. Crossover Operator*

Crossover operator on a pair of chromosomes from parents has done and generate a new pair of chromosomes (offspring). There are several crossover operators like: one-point and two-point crossover. In one-point Crossover, a random position between two genes considered. Then all the genes in the right or left side of the parent moved together to obtained new chromosome.

In two-point crossover, two random positions are randomly choosen and all the genes between these two positions in parent choromoomes replaced. Here one-point crossover has used to generate off springs.

### *3.7. Mutation Operator*

After crossover, mutation operator on chromosome is used. Here is a random selection of two featurs in parent and then changing the places of these two features.

### *3.8. Obtaining New Population*

Then the population generating in mutation and crossover were added to the old population and then multi objective steps (non dominated sorting and crowding distance) are done on new population. And select none dominate solotions, the crowding distance between this answers has changed so multi-objective is performed again.

### *3.9. Stopping Criteria*

When the number of iterations exceeds a maximum number of iterations, the algorithm is terminated. At this point, the best individuals are selected which maximize class discrimination and minimize support vector machine classification error.

## 4. Experimental Results

The prpposed method has been implemented using MATLAB software and tested on some UCI datasets shown in Table 1.

Table 1
Used UCI datasets

| Dataset name | Number of Features | Number of Samples | Number of Classes |
|---|---|---|---|
| Wine | 13 | 178 | 3 |
| Ionosphere | 34 | 351 | 2 |
| Lung-cancer | 56 | 32 | 3 |
| Glass | 9 | 214 | 6 |

In the multi-objective genetic algorithm, population size, crossover and mutation rate are 50, 0.9 and 0.1 respectively. To evaluate the proposed method, linear and kernel based SVM have been co considered where 60% and 40% of each dataset have been utilized as training, and test data sets. Results for multi-objective genetic algorithms have been averaged for best 5 Pareto solutions.

We compared the results with single objective feature selection methods. Table 2 for displaying and comparing one objective method with our method has been adjusted. In Table 2, show the implemented methods and their corresponding abbreviation.

Table 2
Implemented methods and their abbreviation

| **Methods** abbreviation | Methods |
|---|---|
| GA-E | method based on genetic algorithm and SVM classification error |
| GA-MI | method based on genetic algorithm and mutual information |
| GA-LAP | method based on genetic algorithm and Laplacian measure |
| MOGA-EMI | method based on multi objective genetic algorithm for both mutual information and SVM classification error criteria |
| MOGA-ELAP | method based on multi objective genetic algorithm for both Laplacian and SVM classification error criteria |

### *4.1. Comparison Results*

We use multi-objective genetic algorithm to consider filter methods measures and SVM classification error simultaneously. Du to results in

table 3, in most cases considering filter methods criteria along with SVM classification outperforms single-objective methods which use only filter methods measure or only classification error, we have better results for kernel based SVM in most cases .

GAE has better performance than other single objective methods. This shows effectiveness of combination of this criterion with other criteria (mutual information and Laplacian). Table 3 shows that considering Laplacian and mutual information criteria with SVM classification error in multi-objective genetic algorithm provide better results than single-objective approaches.

Between multi-objective proposed methods MOGA-ELAP performance is better than the MOGA-EMI.

## 5. Conclusion

In this paper we have proposed a multi-objective feature selection method which uses Laplacian and mutual information (data-based) measures along with SVM classification error (classification based measure). To this end, we use multi-objective genetic algorithms. Evaluation results on the UCI datasets with medium number of features shows that the proposed method reduces the number of selected features and preserve or improves SVM classification rate. As future works we want to use other evolutionary algorithms and other feature selection criteria like Pearson correlation coefficient, inter and intra class distances.

Table 3

SVM classification accuracy for different multi-objective methods. Best results have identified by bold underlined in the table.

| method | Kernel type (RBF parameters and the degree of the polynomial) | glass | | Lung-cancer | | ionospher | | wine | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | **Number of selected features** | **Accuracy** | **Number of selected features** | **Accuracy** | **Number of selected features** | **Accuracy** | **Number of selected features** |
| GA-E | Linear | 83.72 | 7 | 100 | 24 | 97.13 | 11 | 100 | 8 |
| | RBF(5) | 88.38 | 8 | 100 | 22 | 98.57 | 17 | 100 | 8 |
| | Polnomial (2) | 74.41 | 7 | 100 | 25 | 97.65 | 16 | 100 | 9 |
| GA -MI | Linear | 65.11 | 5 | 83.33 | 21 | 85.71 | 8 | 86.11 | 4 |
| | RBF(5) | 74.41 | 8 | 83.33 | 16 | 90 | 7 | 83.33 | 4 |
| | Polnomial (2) | 62.79 | 6 | 83.33 | 23 | 90 | 7 | 80.55 | 4 |
| GA -LAP | Linear | 74.41 | 4 | 83.33 | 10 | 78.57 | 15 | 52.77 | 1 |
| | RBF(5) | 74.41 | 4 | 83.33 | 22 | 80 | 3 | 41.66 | 1 |
| | Polnomial (2) | 30.23 | 1 | 83.33 | 9 | 72.85 | 4 | 44.44 | 1 |
| MO GA - EMI | Linear | 79.07 | 4 | **100** | 22 | 95.71 | 10 | **100** | 5 |
| | RBF(5) | 79.07 | 7 | 83.33 | 20 | 95.71 | 17 | 97.22 | 6 |
| | Polnomial (2) | **76.74** | 5 | **100** | 20 | **97.86** | 10 | **100** | 8 |
| MO GA - ELAP | Linear | 79.07 | 7 | **100** | 18 | 92.15 | 7 | **100** | 4 |
| | RBF(5) | 79.07 | 8 | **100** | 11 | 95.71 | 12 | **100** | 2 |
| | Polnomial (2) | **88.37** | 4 | **100** | 12 | **97.86** | 20 | **100** | 8 |

## References

[1] R. Porpoladi, "multi-objective feature selection using particle optimization for pattern recognition" master's thesis, Islamic Azad University of Qazvin, 2015. ( In Persian)

[2] F. Long, L. Chia, "Combination of feature selection approaches with SVM in credit scoring" Expert Systems with Applications, vol.37(7), pp. 4902–4909, 2010.

[3] T. Binh, X. Bing, Z. Mengjie, "Simulated Evolution and Learning" Victoria University of Wellington, 2014.

[4] P. Ghamisi, "Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization" IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, vol. 12(2), 2015.

[5] Ch. Yi-Wei, L. Chih-Jen, "Combining SVMs with Various Feature Selection Strategies" Springer Berlin Heidelberg, 2007.

[6] W. Tinghua, Hu. Houkuan, T. Shengfeng, Xu. Jianfeng, "Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels" Expert Systems with Applications, vol.37(9), pp. 6663–6668, 2010.

[7] Y. Jihoon, H. Vasant, "Feature Subset Selection Using A Genetic Algorithm" construction and selection, Springer, 1998.

[8] O. Il-Seok, L. Jin-Seon, M. Byung-Ro, "Hybrid Genetic Algorithms for Feature Selection" Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26(11), PP. 1424-1437, 2004.

[9] N. Omar, F. Jusoh, F,R. Ibrahim, MS. Othman, "Review of Feature Selection for Solving Classification Problems" Journal of Information System Research and Innovation, vol.3, PP. 64-70, 2013.

[10] T. M. Cover, J. A. Thomas, "Information Theory and Statistics", pp. 279-335, 2001.

[11] J. R. Vergara, P. A. Este´vez, "A review of feature selection methods based on mutual information" Neural Comput & Applic, vol. 24, pp. 175–186, 2014.

[12] L. Zhu, L. Miao, D. Zhang, "Iterative Laplacian Score For feature Selection", Springer-Verlag Berlin Heidelberg, pp. 80-87, 2012.

[13] M. Belkin, P. Laplacian Niyogi, eigenmaps for dimensionality reduction and data representation. Neural compute, 2003.

[14] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A Fast and Elitist Multi objective Genetic Algorithm NSGA-II, IEEE Transactions on evolutionary computation, pp. 182-197, 2002.