# Feature selection based on filter approach using mutual information with Genetic algorithm

S.Sivakumar
Research Scholar
Department of Computer Science
Periyar University
Salem, India-636 011.

Dr.C.Chandrasekar
Associate Professor
Department of Computer Science
Periyar University
Salem, India-636 011

***Abstract*— Feature selection is a well-known and important problem in pattern recognition, data mining and information retrieval. Methods for feature selection are typically classified as filter or wrapper. Filter methods do not depend on any underlying learning algorithm and have a wider applicability while wrapper methods rely heavily on the specific structure of learning algorithms. This paper proposes a new feature selection method using a mutual information based criterion that measures the importance of a feature and optimize the selected features via genetic algorithm. In order to evaluate the significance of the algorithm, the LIDC-IDRI database images are used as the dataset for the experiment.**

***Keywords-: feature selection approaches, mutual information, genetic operators, classification.***

## I. INTRODUCTION

Classification problems often have a large number of features, but not all of them are useful for classification. Irrelevant and redundant features may even reduce the classification accuracy. Feature selection is a process of selecting a subset of relevant features, which can decrease the dimensionality, shorten the running time, and/or improve the classification accuracy. There are two types of feature selection approaches, i.e. wrapper and filter approaches. Their main difference is that wrappers use a classification algorithm to evaluate the goodness of the features during the feature selection process while filters are independent of any classification algorithm. Feature selection is a difficult task because of feature interactions and the large search space. Therefore, feature selection is proposed to increase the quality of the feature space, reduce the number of features and/or improve the classification performance [2][3][4]. Feature selection aims to select a subset of relevant features that are necessary and sufficient to describe the target concept [1]. By reducing the irrelevant and redundant features, feature selection could decrease the dimensionality, reduce the amount of data needed for the learning process, shorten the running time, simplify the structure and/or improve the performance of the learnt classifiers [1]. Naturally, an optimal feature subset is the smallest feature subset that can obtain the optimal performance, which makes feature selection a multi-objective problem [5]. Note that feature selection algorithms choose a subset of features from the original feature set, and do not create new features.

Existing feature selection methods can be broadly classified into two categories: filter approaches and wrapper approaches [1][2]. Wrapper approaches include a classification algorithm as a part of the evaluation function to determine the goodness of the selected feature subsets. Filter approaches use statistical characteristics of the data for evaluation and the feature selection search process is independent of any classification algorithm. Filter approaches are computationally less expensive and more general than wrapper approaches while wrappers are better than filters in terms of the classification performance [1]. In this paper we aimed to select the optimal number of features using filter approach with the help of genetic algorithm via mutual information.

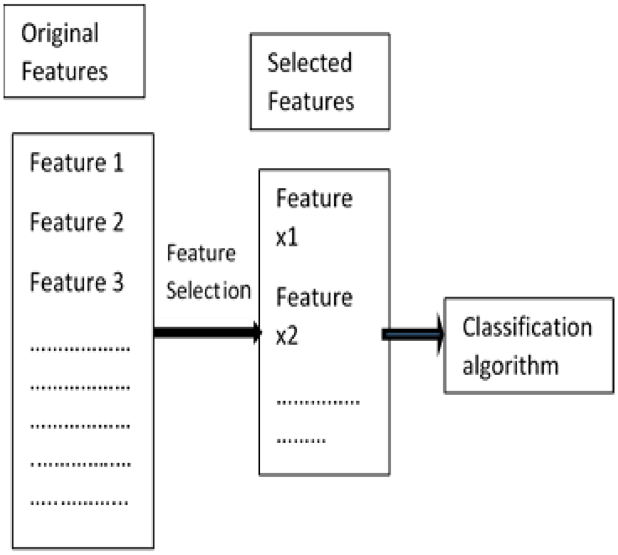

Figure1: Overview of Feature Selection process

## II. FILTER BASED FEATURE SELECTION

Figure 1 shows the diagram of a feature selection system taking a filter algorithm. In filter algorithms, the search process is independent of any classification algorithm. The goodness of feature subsets are evaluated based on a certain criterion like distance measure, information measure and consistency measure [1]. Information theory provides a way to measure the information of the random variables [6]. Information theory can be viewed as a branch of mathematics and it is also related to electrical engineering, bioinformatics, and computer science [7]. Information theory provides different ways to quantify uncertainty.

## III. Mutual Information

The entropy is a measure of the uncertainty of random variables. Let X be a random variable with discrete values, its uncertainty can be measured by entropy H(X), which is defined as

$$H(X) = -\sum_{x\in X} p(x) log2p(x)$$

where p(x) = Pr(X = x) is the probability density function of X. Note that entropy does not depend on actual values, just the probability distribution of the random variable. For two discrete random variables X and Y with their probability density function p(x; y), the joint entropy H(X; Y) is defined as

$$H(X,Y) = -\sum_{x\in X, y\in Y} p(x,y) log2\, p(x,y)$$

The information shared between two random variables is defined as mutual information. Given variable X, how much information one can gain about variable Y , which is mutual information I(X; Y ).

$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= -\sum_{x\in X, y\in Y} p(x,y) log2 \frac{p(x,y)}{p(x)p(y)}$$

According to above equation, the mutual information I(X; Y) will be large if two variables X and Y are closely related. Otherwise, I(X; Y) = 0 if X and Y are totally unrelated.

## IV. Genetic Algorithm

Genetic algorithms are adaptive algorithms for finding the global optimum solution for an optimization problem. The canonical genetic algorithm developed by Holland is characterized by binary representation of individual solutions, simple problem-independent crossover and mutation operators, and a proportional selection rule [12].

GAscomprise a subset of these evolution-based optimization techniques focusing on the application of selection, mutation, and recombination to a population of competing problem solutions. In our GA-based feature subset selection, each individual is represented as a binary string encoding a feature subset. If the data consist of N features, an individual will be an N-bit binary string. If a bit is 1 the feature is chosen in the feature subset; if 0 it is not. Each individual in the population is thus a candidate feature subset [9-12]. The following are the steps involved in GA based feature selection.

(1) Generating Initial Population:

In the initialization phase, the first thing to do is to decide the coding structure. Coding for a solution, termed a chromosome in GA literature, is usually described as a string of symbols from {0, 1}. These components of the chromosome are then labeled as genes. The number of bits that must be used to describe the parameters is problem dependent. Let each solution in the population of m such solutions xi, i=1, 2,.,. m, be a string of symbols {0, 1} of length N, because number of feature is N.

(2) Evaluate the fitness:

In order to evaluate the fitness of the initial population, calculate the mutual information between the feature subset and the class variable. If the fitness value is satisfied means terminate and produce the result, otherwise follow the next steps.

(3) Selection process:

GA uses proportional selection, the population of the next generation is determined by *k* independent random experiments; the probability that individual xi is selected from the tuple ($x_1$, $x_2$, . . . , $x_m$) to be a member of the next generation at each experiment is given by

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{m} f(x_j)} > 0$$

This process is also called roulette wheel parent selection and may be viewed as a roulette wheel where each member of the population is represented by a slice that is directly proportional to the member's fitness. A selection step is then a spin of the wheel, which in the long run tends to eliminate the least fit population members.

(4) Crossover:

Crossover is an important random operator in GA and the function of the crossover operator is to generate new or 'child' chromosomes from two 'parent' chromosomes by combining the information extracted from the parents. The method of crossover used in GA is the one-point crossover as shown in Figure 3. By this method, for a chromosome of a length N, a random number c between 1 and N is first generated. The first child chromosome is formed by appending the last N−c elements of the first parent chromosome to the first c elements of the second parent chromosome. The second child chromosome is formed by appending the last N−c elements of the second parent chromosome to the first c elements of the first parent chromosome.

**Parent1: 1 0 1 0 || 0 0 1 1 0 1 → child1: 0 1 1 0 0 0 1 1 0 1**
**Parent2: 0 1 1 0 || 1 1 0 1 0 1 → child2: 1 0 1 0 1 1 0 1 0 1**

Figure 2: crossover operation between parent1 and parent2

(5) Mutation: Mutation is another important component in GA. It operates independently on each individual by probabilistically perturbing each bit string. A usual way to mutate used in CGA is to generate random number v between 1 and l and then make a random change in the *v*th element of the string with probability $P_m^{\in}(0, 1)$, which is shown in Figure 3.

**Parent: 1 0 1 0 1 1 0 1 0 1 →Mutation→ 1 0 1 0 1 0 0 1 0 1**

Figure 3: Mutation operation

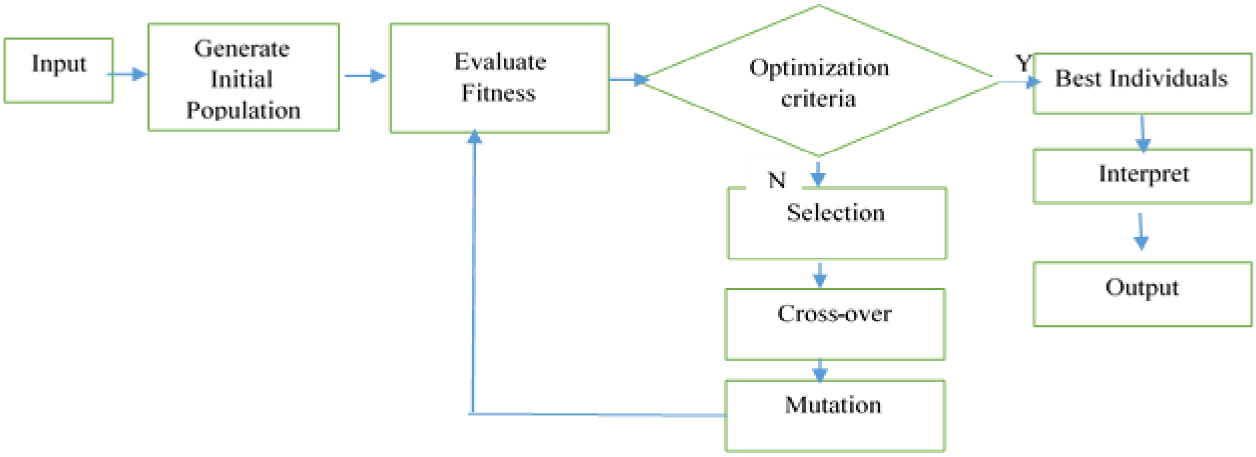

Figure 4: Genetic Algorithm process flow chart

After mutation the newly generated child population will be evaluated against the fitness value, if its fail repeat the steps (3) to (5) until its reach maximum number of generations.

## V. Feature Selection using Mutual Information with GA

Information theory, mainly mutual information, has been applied to filter feature selection to measure the relationship between the selected features and the class labels. Since mutual information are capable to evaluate the relationship between variables, they have been applied to feature selection to measure the relationship between the selected features and the class labels. In the following formulation *F* refers to a set of features and *C* to the class labels [3].

$$I(F,C)=\int\int p(f,c)\log\frac{p(f,c)}{p(f)p(c)}dfdc$$

Some approaches evaluate the mutual information between a single feature and the class label. This measure is not a problem. The difficulties arise when evaluating entire feature sets. The necessity for evaluating entire feature sets in a multivariate way is due to the possible interactions among features [4]. While two single features might not provide enough information about the class, the combination of both of them could, in some cases, provide significant information. For the mutual information between *N* variables $X_1$, $X_2$ ... $X_N$, and the variable $Y$, the chain rule is[5]:

$$I(X1,X2,\dots,XN;Y)=\sum_{i=1}^{N} I(Xi;Y\,|Xi-1,Xi-2,\dots\dots,X1)$$

The usual approach for calculating mutual information is to measure entropy and substitute it in the mutual information formula. Mutual information is considered to be a suitable criterion for feature selection.

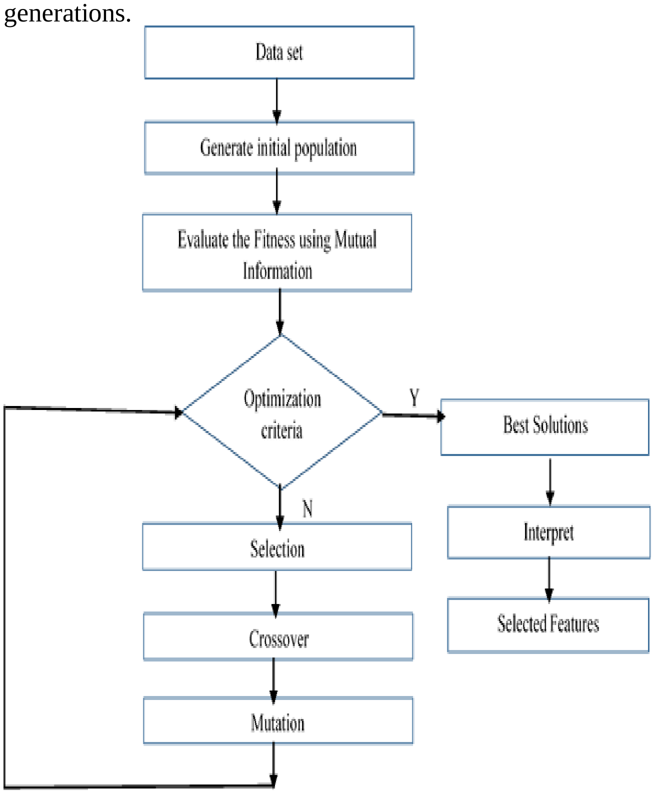

Figure 5: Process flow for FS using MI with GA

## VI. Support Vector Machines

Support vector machines (SVMs) are a popular machine learning method. They are based on the concept of decision planes that define decision boundaries. The main idea of SVMs is to use a kernel function to map the input data to a higher-dimensional space, where the instances are linearly separable. In the high-dimensional space, SVMs construct a hyper plane or a set of hyper planes, which are used to create decision boundaries for classification [13]. SVMs are inherently two-class classified [8]. Each hyper plane is expected to separate between a set of instances having two classes. Instances are classified based on what side of these hyper planes they fall on. SVMs aim to maximise the distances between the hyper planes and the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [13]. A particular advantage of SVMs over other learning algorithms is that they are based on sound mathematics theory and can be analyzed theoretically using concepts from the computational learning theory [14].

## VII. Experimental Results and Discussion

In order to evaluate the performance of the modified PSO for Feature Selection the LIDC-IDRI Lung CT scan images were used. The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions. It is a web-accessible international resource for development, training, and evaluation of computer-assisted diagnostic (CAD) methods for lung cancer detection and diagnosis. Each study in the dataset consist of collection of slices and each slice of the size of 512 X 512 in DICOM format. The lungs image data, nodule size list and annotated XML file documentations can be downloaded from the National Cancer Institute website [16]. For the experiment we taken 175 Non-Cancer Lung CT scan images and 225 Cancer Lung CT images from the LIDC dataset. All the CT scan images are preprocessed through wiener filter and the lung portion is extracted through morphological operations. From the segmented lung portion, both the first order statistical features namely mean, variance, standard deviation, skewness, and kurtosis and second order statistical features namely GLCM based 14 Haralick features, GLRLM based 7 features, GLDM based 8 features and GLRLM based 7 features are calculated. These calculated features are used as the dataset for the proposed method.

Table 1: Parameters used in the proposed work

| Parameters | Value |
|---|---|
| Population Size | 100 |
| String Length | 5, 13, 7, 8, 7 |
| Number of Generations | 300 |
| Probability of Crossover | 0.95 |
| Probability of Mutation | 0.01 |
| Elite Count | 2 |
| Type of Mutation | Uniform |
| Type of Selection | Roulette-wheel |

Table 2: List of selected features using the proposed method

| Run | First Order Features (5) | GLCM (14) | GLRLM(7) | GLDM(8) | GLGLM(7) |
|---|---|---|---|---|---|
| 1 | [1,2] | [1,2,3,5,8,9,10,11] | [1,2,6,7] | [1,3,5,6,8] | [1,2,6,7] |
| 2 | [2,3,5] | [1,2,4,6,9,10] | [1,2,4,7] | [1,3,5,6,7] | [1,3,4,5] |
| 3 | [2,4] | [1,2,5,7,11] | [1,2,5,7] | [1,2,5,6,8] | [1,4,5,6,7] |
| 4 | [1,4] | [1,2,4,5,6,8,10,13] | [1,2,4,5] | [3,5,6,7] | [1,3,5,6] |
| 5 | [1,2,4] | [1,3,4,8,9,11,12] | [1,2,5,6,7] | [1,3,6,7] | [3,4,7] |
| 6 | [1,2,4] | [1,3,7,9,10,12,13] | [1,3,5] | [1,5,6] | [2,3,5,6] |
| 7 | [1,4,5] | [1,4,5,7,8,10,12] | [2,5,7] | [3,4,5,7,8] | [1,4,6,7] |
| 8 | [1,3,5] | [1,2,7,8,10,13] | [1,3,5,7] | [3,4,5,6] | [1,2,4,6,7] |
| 9 | [1,5] | [1,2,5,8,9,10,12] | [1,2,4,6] | [1,2,5,6, 8] | [3,4,6,7] |
| 10 | [1,2,4] | [1,3,4,5,7,8,11,13] | [1,2,6] | [1,3,4,6,7] | [2,3,4,6] |

Table 1 shows the parameters used for feature selection using mutual information with genetic algorithm and the table 2 shows the list of selected features by that algorithm. In the experiment, the instances in each dataset are randomly divided into two sets: 70% as the training set and 30% as the test set. From table1, the modified PSO Feature selection yields better accuracy with minimal set of features.

Table 3: Classification accuracy of Selected Features using different methods evaluated by SVM classifier

| Dataset | Unreduced feature set | | Selected Features using Mutual Information | | Selected Features using MI with GA | |
|---|---|---|---|---|---|---|
| | Number of features | Classification accuracy (%) | Number of features | Classification accuracy (%) | Number of features | Classification accuracy (%) |
| First Order | 5 | 73 | 4 | 81 | 3 | 88 |
| GLCM | 14 | 78 | 11 | 83 | 9 | 94 |
| GLRLM | 7 | 67 | 5 | 80 | 4 | 91 |
| GLDM | 8 | 69 | 6 | 78 | 5 | 89 |
| GLGLM | 7 | 76 | 6 | 91 | 4 | 96 |

From the table 3, feature selected by the mutual information with the genetic algorithm yields the good classification accuracy where compare with the unreduced feature set and the features selected by the mutual information oriented approach.

## CONCLUSION

In this work, mutual information based feature selection using genetic approach is used to select the feature subset for the classification purpose. From the result, the classification accuracy of the SVM classifier for the proposed method performs significantly superior where compare with the basic mutual information based feature selection approach. The MI with GA based feature selection approach to be seen that, reducing the number of features by selecting only the significant features and improve the classification results in the form of classification accuracy.

## ACKNOWLEDGMENT

The First Author extends his gratitude to UGC as this research work was supported by Basic Scientific Research (BSR) Non-SAP Scheme, under grant reference number, F-41/2006(BSR)/11-142/2010(BSR) UGC XI Plan.

## REFERENCES

[1] M. Dash and H. Liu, “Feature selection for classification,” Intelligent Data Analysis, vol. 1, no. 4, pp. 131–156, 1997.

[2] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, pp. 273–324, 1997.

[3] H. Liu and H. Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[4] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[5] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, “Feature selection using multi-objective genetic algorithms for handwritten digit recognition,” in 16th International Conference on Pattern Recognition (ICPR’02), vol. 1, pp. 568– 571, 2002.

[6] C. Shannon and W. Weaver, “The Mathematical Theory of Communication.” Urbana: The University of Illinois Press, 1949.

[7] T. M. Cover and J. A. Thomas, Elements of information theory. New York, NY, USA: Wiley-Interscience, 1991.

[8] S.Sivakumar and C.Chandrasekar, “Lung Nodule Detection UsingFuzzy Clustering and Support Vector Machines”, International Journal of Engineering and Technology, vol. 5, no. 1, pp. 179-185, 2013.

[9] Mitchell T, “Machine Learning”, McGraw Hill, New York, 1997.

[10] Duda R, Hart P, and Stork D, “Pattern Classification”, Wiley-Interscience, New York, 2000.

[11] Bishop C, “Pattern Recognition and Machine Learning”, Springer, New York, 2006.

[12] Siedlecki W and Sklansky J, “A note on genetic algorithms for large-scale feature selection,” Pattern Recognition Letters, Vol. 10, pp.335-347, 1989.

[13] C. W. Hsu, C. C. Chang, and C. J. Lin, “A practical guide to support vector classification,” 2003.

[14] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, “Support vector machines,” IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18 –28, 1998.

[15] S.Sivakumar and C.Chandrasekar, “Lung Nodule Segmentation through Unsupervised Clustering Models”, Procedia Engineering, vol.38, pp. 3064-3073.

[16] S.Sivakumar and C.Chandrasekar, “A Study on Image Denoising for Lung CT Scan Images”, International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), vol. 7, no. 1, pp.86-91.