ISSN:0975-9646

Feature Selection Using Genetic Algorithm with Mutual Information

S. Sivakumar, Dr.C.Chandrasekar

Department of Computer Science, Periyar University, Salem-11, Tamilnadu, India

Abstract:Feature selection is the problem of selecting a subset of features without reducing the accuracy of representing the original set of features. It is the most important step that affects the performance of a pattern recognition system. In this paper, genetic algorithm (GA) is used to implement a feature selection in filter based method, and the mutual information is served as a fitness function of GA and k-NN is used to evaluate the accuracy of the selected feature. The proposed feature selection method is applied to the features extracted from the Lung CT scan images. Experimental results shows that proposed feature selection method simplifies features effectively and obtains a higher classification accuracy compared to the unreduced dataset classification accuracy.

Keywords: ***Feature Selection, Genetic Algorithm, Mutual Information, Classification accuracy***

I. INTRODUCTION

The main purpose of feature subset selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy. There are two major approaches to dimensionality reduction: feature selection and feature transform. Whilst feature selection reduces the feature set by discarding the features which are not useful for some definite purpose (generally for classification), feature transform methods (also called feature extraction) build a new feature space from the original variables. Less discriminatory features are eliminated, leaving a subset of the original features which retains sufficient information to discriminate well among classes [1]. For classical pattern recognition techniques, the patterns are generally represented as a vector of feature values. The selection of features can have a considerable impact on the effectiveness of the resulting classification algorithm. Consider a feature set, $F = \{f_0; f_1; \ldots; f_N\}$. If f_0 and f_1 are dependent, that is they always move together, then one of these could be discarded and the classifier has no less information to work with. This has the benefit that computational complexity is reduced as there is smaller number of inputs. Often, a secondary benefit found is that the accuracy of the classifier increases. This implies that the removed features were not adding any useful information but they were also actively hindering the recognition process. Feature selection is a field with increasing interest in machine learning. The literature differentiates among three kinds of feature selection: Filter method, wrapper method and on-line. Filter feature selection does not take into account the properties of the classifier, as it performs statistical tests to the variables, while wrapper feature selection tests different feature sets by building the classifier. Finally, on-line feature selection incrementally adds or removes new features during the learning process. All of these methods are based on some feature selection criterion, for example, the criterion of wrappers is the classification performance while the criterion of filters usually is some statistical test on the variables [2] [3] [4].

II. MUTUAL INFORMATION (MI)

When there are thousands of features, wrapper approaches become infeasible because the evaluation of large feature sets is computationally expensive. Filter approaches evaluate feature subsets via different statistical measures. Among the filter approaches, a fast way to evaluate individual features is given by their relevance to the classification, by maximizing the mutual information between each single variable and the classification output. In this work we use the mutual information criterion and we estimate its value directly from the data. This kind of estimation methods bypass the estimation of the distribution of the samples. Thus, the low number of samples in a high dimensionality is not a problem anymore. Information theory offers a solid theoretical framework for many different machine learning problems. In the case of feature selection, information theoretic methods are usually applied in the filter feature selection way. A classical use of information theory is found in several feature ranking measures. These consist in statistics from the data which score each feature *Fi* depending on its relation with the classes. One of the most relevant contributions of information theory to the feature selection research is the use of mutual information for feature evaluation. In the following formulation *F* refers to a set of features and *C* to the class labels [6].

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

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 [7]. 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]:

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

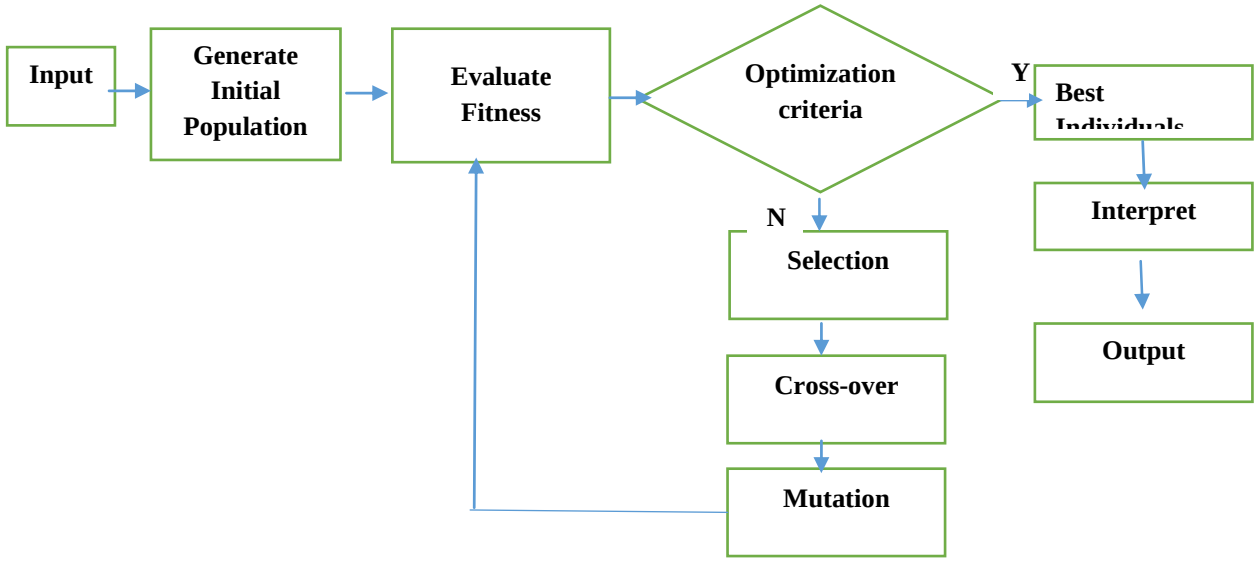

Figure 1: Genetic Algorithm process flow chart

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. Mutual information is a measure of the reduction of uncertainty about the class labels, due to the knowledge of the features of a data set [8].

III. GENETIC ALGORITHM (GA)

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].
GAs comprise 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 2. 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 a random number v between 1 and l and then make a random change in the v^{th} element of the string with probability $P_m \epsilon (0, 1)$, 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

(6) After mutation the newly generated child population will be evaluated against the fitness value, if its fail repeat the steps (3) to (6) until its reach maximum number of generations.

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of the GA based feature selection with mutual information, the LIDC-IDRI Lung CT scan images were used as a dataset. 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. For the experiment we taken 170 Non-Cancer Lung CT scan images and 340 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 (mean, variance, standard deviation, skewness, and kurtosis) and second order statistical features including GLCM based 14 Haralick features and GLRLM based 7 features are extracted. These features are taken as the input for the GA based feature selection with mutual information. In order to evaluate the selected features, the k-NN Classification model is used.

Table 1: Parameters used in the proposed GA with MI Feature selection

Parameters	**Value**
Population Size	100
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

From the table 2, the three different type of features which are extracted from the Lung CT scan images namely first order statistical features, GLCM based Haralick features and GLRLM based features used in our experiment. The three features sets have different number of features (5, 14, 7), with two classes and 510 instances as the representative samples of the problems that the proposed algorithms can address. In the experiments, the instances in each dataset are randomly divided into two sets: 70% as the training set and 30% as the test set with parameters from the table 1. From table2, GA with MI yields better classification accuracy with minimal set of features where compare with unreduced feature set and the mutual information of the selected features are also high.

V. CONCLUSION

In this paper, we have proposed Genetic Algorithm with Mutual Information for feature selection to overcome the limitations of classification of survival analysis in lung cancer. Genetic Algorithm with Mutual Information feature selection is capable of searching the optimal features for survival classification. The use of k-NN classifier alone does not improve the average classification accuracy. Genetic Algorithm with Mutual Information feature selection with k-NN is far surpassed the efficiency of classification result. From the result, the classification accuracy for k-NN classifier with Genetic Algorithm with Mutual Information feature selection performs significantly superior to the k-NN classifier without feature selection. It could be seen that reducing the number of features by selecting only the significant one improved the classification accuracy. Based on the experimental result, it may appropriate to suggest feature selection for solving classification problem for survival analysis in lung cancer.

Table 2: Performance Analysis of Ga with MI Based Feature Selection

Feature Set	**Unreduced Feature set**		**Reduced Feature set using GA with MI**		
	MI value for the Features set	Classification accuracy	Number of Selected Features	MI value for the Selected Features	Classification accuracy
First Order Statistical Features (5)	0.7148	76.52%	3	0.7763	84.42%
GLCM based Features (14)	0.6814	73.41%	8	0.7581	81.86%
GLRLM based Features (7)	0.7642	79.28%	4	0.8143	90.48%

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] Lee K, Joo J, Yang J, and Honavar V, "Experimental comparison of feature subset selection using a GA and ACO algorithm," in Proceedings of the 2nd International Conference on Advanced Data Mining and Applications, pp. 465-472, 2006.

[2] Escolano F, Suau P and Bonev B, "Information Theory in Computer Vision and Pattern Recognition", Springer, Computer Imaging, Vision, Pattern Recognition and Graphics, New York, 2009.

[3] Cover. M and Thomas. J, "Elements of Information Theory", Wiley Interscience.

[4] Neemuchwala. H., Hero. A., and Carson P, "Image registration methods in high-dimensional space", International Journal on Imaging, 2006.

[5] Vasconcelos, N. and Vasconcelos, M, "Scalable discriminant feature selection for image retrieval and Recognition", Computer Vision and Pattern Recognition Conference (CVPR04) proceedings, pp. 770-775.

[6] Guyon I and Elissee A, "An introduction to variable and feature selection", Journal of Machine Learning Research, volume 3, pp. 1157-1182.

[7] Peng H, Long F and Ding C, "Feature selection based on mutual information: Criteria of max-dependency, maxrelevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 27, pp. 1226-1238, 2005.

[8] Kozachenko L. F. and Leonenko N. N, "Sample estimate of the entropy of a random vector", Problems Information Transmission, 23(1):95-101, 1987.

[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.