Improvement of Feature Selection in multi-phase CT images of hepatic lesions

Shuqin Wang, Yan Sun, Qionghua Weng, Jun Ye, Lixu Gu*, Lijun Qian and Jianrong Xu

Abstract—**With the development of the image processing technology and artificial intelligence, feature selection techniques have been widely used in various fields; it can help us to identify important and irrelevant (unimportant) features. In this study, a new refined feature selection module which utilizes two-step selection method was proposed in computer-aided diagnosis (CAD) system for liver disease. It is based on filter and wrapper method, Support Vector Machine (SVM) and Genetic Algorithm (GA). Our contribution is to propose an approach that shows great advantage in the ability of accommodating multi feature selection search strategies and combining filter and wrapper method, especially in identifying optimal and minimal feature subsets for building the classifier. Experimental results show that the algorithm proposed in this paper can find feature subsets with smaller size and higher classification accuracy.**

Index Terms—Feature selection, Computer-aided diagnosis, genetic algorithm, support vector machine

I. INTRODUCTION

The number of liver cancers diagnosed in the US and throughout the world is increasing at an alarming rate and the number will continue to increase over the next few decades [1]. Liver cancer is a lethal cancer with untreated patients rarely surviving more than one year. Early detection and treatment is the most useful way to reduce cancer deaths. With the artificial intelligence techniques and imaging technology advanced, the interpretation of medical images has been greatly enhanced, which contributed to early diagnosis. However, it also leads to a huge amount of feature data.

For image recognition, more inputted characteristic items do not means better, they may produce a lot of false positive findings. “Information overload” will weaken the classification performance. In addition, when inputted features are increased, the training samples required for classification will grow in exponential. Therefore, in the liver diagnose system, how to choose the right feature set which contribute to the high classification accuracy from a number of features obtained by feature extraction method such as first order statistics (FOS), Gray Level Co-occurrence Matrix (GLCM) [2] or temporal method [3] is the key issue.

The basic task of feature selection is how to find out the most effective feature subset from high-dimensional features. It includes the following two sub-problems: 1) search strategy 2) the issue of evaluation functions. Based on the search strategy used in feature selection algorithm, it can be categorized into 3 classes: (1) global optimal search strategy (2) Randomized algorithms (Such as Genetic Algorithm (GA) [4].): The method is a searching method for solving in Local minima which adds randomly. (3) Sequential algorithms: The algorithm is added or subtracted features sequentially, Such as Sequential Forward Selection (SFS), Sequential Backward Elimination (SBE)[5], Plus l Take-Away r Selection (PTA), Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Elimination (SFBE) [6][7]etc. On the other hand, there are two types of feature selection framework, derived from the nature of the evaluation function $J(\cdot)$ used: filters [8] and wrappers [9]. In a filter framework, $J(\cdot)$ measures the performance of a feature set in a manner that does not include the classification algorithm which will eventually use the features. In a wrapper framework, $J(\cdot)$ incorporates the classification algorithm.

In this paper, a new method that utilizes the two-step selection approach was proposed to choose the most relevant features from a large feature set. This two-step selection method can be described as: firstly, apply traditional sequential algorithms such as SFS, SBE, SFFS, SFBE, PTA to obtain five different feature subsets which will be used to generate a new feature set and then utilizes GA to search feature space from the new feature group by the fitness function designed by the accuracy of SVM [10]. The advantages of this approach include the ability to accommodate different feature selection search strategies and combine filter and wrapper method, which makes the system can find a small optimal feature subsets that perform well for a particular inductive learning algorithm of interest to build the classifier.

The rest of this paper is organized as follows. Section II introduces the algorithm; feature selection method will be expounded. In section III, we describe the experimental results. Section IV presents conclusions and discussions.

This paper will be presented by the third author Qionghua Weng。

Shuqin Wang，School of Software, Shanghai Jiaotong University P.R.China (e-mail: hyphood@ yahoo.com.cn；phone：86-13564280748).

Yan Sun , School of Software, Shanghai Jiaotong University P.R.China (e-mail: sunyan@ cs.sjtu.edu.cn)

Qionghua Weng, Jun Ye, School of Software, Shanghai Jiaotong University P.R.China.

Lixu Gu, Med-X Research Institute, Shanghai Jiaotong University, Shanghai, P.R.China(corresponding author, e-mail: gulixu@sjtu.edu.cn)

Lijun Qian, Jianrong Xu, Department of Radiology,Renji Hospital, SJTU, Shanghai, P.R.China

II. Algorithm

Here, we describe the specific meaning of some concepts we used.

Feature Set/Feature Vector: This computer-aided diagnosis system (CAD) includes four modules: Regions of Interest (ROIs) extraction, feature extraction, feature selection and feature classification. The CAD system for characterization of hepatic lesions from Multi-Phase computed tomography (CT) images detects kinds of pathologies in ROI which were drawn by an experienced radiologist are categorized into 4 classes: normal, cyst, haemangioma and carcinoma. Texture features extracted from the multi-phase liver images were based on the First-Order Statics, GLCM and temporal method. Consequently，a total of 48 texture features have been received which were the components of the feature set.

***Criterion*:** One type of filter measures will be discussed: inter-class distance. Theoretically, the lager the separation between classes is, the easier it will be to define a decision boundary, and to achieve a higher recognition rate on novel data. Methods such as the Mahalanobis distance, see Devijer and Kittler [5] and Duda and Hart[11] have already be used. Mahalanobis distance will be used for the experiments of this work. Mahalanobis distance can be defined as dissimilarity measure between two random vectors $\vec{X}$ and $\vec{Y}$ of the same distribution with the covariance matrix S:

$$d\left(\overline{X},\overline{Y}\right)=\sqrt{\left(\overline{X}-\overline{Y}\right)^{T}S^{-1}\left(\overline{X}-\overline{Y}\right)}$$

A. Sequential feature selection method

The most straightforward approach to the feature selection problem can be described as follows: 1. Examine all possible subset of size m of the original feature set; 2. select the subset with the largest value of $J(\cdot)$. As previously mentioned, Mahalanobis distance was employed in five different search strategies: SFS, SBE, SFFS, SFBE, and PTA.

SFS&SBE: The most common searches are SFS and SBE. SFS starts with an empty feature set and has an inclusion operator, adding one feature at a time, attempting to maximize $J(\cdot)$. On the contrary SBE starts with a set of all the available features, and uses an exclusion operator only. See Devijer and Kittler [5].

SFFS&SFBE&PTA: More sophisticated techniques are the Plus l Take Away r and the Sequential Floating Search that may operate either forward (SFFS) or backward (SFBE) [6]. These methods allow backtrack when they find improvement compared to previous feature sets in same size. PTA use l steps of SFS and then r steps of SBE, while SFFS and SFBE methods allow l and r floating i.e. change at each step [7].

In the problem addressed here, the algorithm described before have been used to obtain five different feature subsets and then we combined these feature subsets to one new feature vector, which is the input of the next step feature selection.

B. Genetic Algorithm

A genetic algorithm (GA) is a stochastic search approach, inspired by evolutionary biology such as inheritance, mutation, selection and crossover. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated; multiple individuals are selected stochastically from the current population based on their fitness, and recombined or mutated to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. According to [4], [12], GA has already been applied to a different field of the feature selection problem.

C. Support Vector Machine

According to [13], the SVM has been extensively used as a classification tool with a great deal of success in a variety of areas from object recognition [14] to classification of cancer morphologies. Unlike other machine learning methods (such as neural network), it's not easy to over-fitting. SVM has showed unique advantages and good prospects in solving the problems of small sample, nonlinear, high-dimensional recognition. The aim of SVM is to devise a computationally efficient way of learning 'good' separating hyperplanes in a high dimensional feature space [10]. That is to find the maximal margin hyperplane in an appropriately chosen kernel-induced feature space. The feature space can be explicitly computed for a kernel.

In support vector machine, the choice of kernel function is important. By selecting the kernel function, a sample can be mapped to a high-dimensional space where the optimal separating hyperplane can be constructed in. Choose a different kernel function is equivalent to choose a different inner product, which means use different standards to evaluate the degree of similarity. At present, there are four kinds of the most commonly used kernel function as follows:

1) Linear：$K(x,y)=x^{T}\cdot y$;

2) Polynomial:

$K(x,y)=\left(x\cdot y+1\right)^{d}$, d is the power;

3) Radial Basis Function：

$K(x,y)=\exp\left\{-\gamma\left\|x-y\right\|^{2}\right\}$;

4) Sigmoid：$K(x,y)=\tan\left\{w\left(x\cdot y\right)+c\right\}$.

According to the characteristics between texture features of liver, this study used Radial Basis Function (RBF) kernel which will allow a support vector to have a strong influence over a larger area. The RBF kernel mapped samples to a higher-dimensional space, which can deal with the relationship of the non-linear between category labels and attributes. The linear kernel is a special case of RBF. Therefore, RBF kernel can be applied to any distribution of the samples through appropriate choice of parameters.

D. Algorithm Realization

The multi-phase abdominal CT imaging are performed by using a 16-row CT scanner (Bright Speed, GE Medical Systems, Milwaukee, WI, USA) with a spatial resolution of 512x512 pixels and 16-bit gray-level. After registration, under the guidance of experts to outline the ROIs, extracting a total of 48 feature parameters based on FOS, GLCM and temporal method. To the FOS, there are 7 features correspond to mean, variance, central moment, kurtosis and skewness etc; to the GLCM, there are 32 features correspond to angular second moment, correlation, contrast, homogeneity, cluster tendency, entropy and depend on inter-sample spacing and angular directions; to temporal method, there are 9 features relies on the vascular structures, differences in vessel growth between different hepatic diseases' lesions can potentially be characterized the contrast uptake and washout of the tissue, correspond to relative signal Intensity[3], intensity change tendency[3], signal enhancement ratio[3].

Our goal is to use the two-step feature selection algorithm proposed in the paper based on filter and wrapper method to extract the optimal feature subset from all 48 features, improve the classification accuracy. The algorithm processes was shown in Figure 1 and Figure 2. Figure 1 is the first step in proposed feature selection method, the Mahalanobis distance criterion employed to the five different search strategies respectively and then five different feature subsets would be obtained. Finally, these feature vectors were merged with each other to generate the new feature set. Figure 2 is the second step of the algorithm: GA was applied to search the feature space; meanwhile, SVM was utilized to train the full feature vector and design the fitness function.

1) Chromosome encoding: In the applied GA-based feature selection techniques, coding is the primary problem. Binary encoding scheme was use to the issue. The feature subset is represented by a binary string composed with zero or one character, which called an individual. Each individual includes one chromosome. Zero or one character indicates the absence or presence of a feature. In this paper, a total of 48 texture features, so the length of the individual is 48.

2) Fitness function design: The aim of the study was to obtain the best classification accuracy, thus, we applied the accuracy of SVM to design the fitness function. The fitness function should be designed to meet the requirements of the higher accuracy rate, the small number of features and easier to calculate. To achieve this goal, this study introduced formula (1) to build the fitness function, that is, let the fitness function proportional to classification accuracy, inversely proportional to the number of characteristics. The function formula as follows:

$$f = w_a \times P + w_f / d \qquad (1)$$

Where, f is fitness, the higher the fitness value of chromosomes the greater probability of the individuals survived to the next generation; w_a is the weight of classification accuracy; w_f is the weight of the number of selected features; P is the classification accuracy. The value of w_a and w_f can be adjusted according to actual needs, commonly, the value of w_a can choose between 0.75 ~ 1.In this study, w_a choose 0.8, w_f choose 0.2. The formula (1) tells that in the condition of the same number of features, the higher the classification accuracy the higher the fitness value.

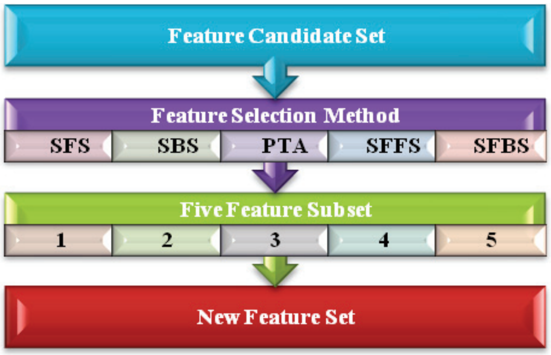


Figure1. First Step Generate the New Feature Set

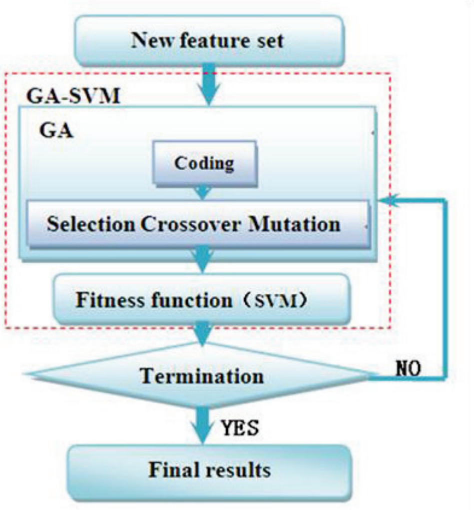


Figure2. Second Step Generate the Final Results

III. Results and Discussion

In the experiment, a total of 179 multi-phase abdominal CT images, from both patients and healthy controls, were acquired from RenJi hospital, Subsidiary of Shanghai Jiaotong University, using a 16-row CT scanner (BrightSpeed, GE Medical Systems, Milwaukee, WI, USA) with a spatial resolution of 512x512 pixels and 16-bit gray-level. The parameters of CT set are as follows: 120 kVp and 140 to 250 mAs, 0.8-second tube rotation time, 16 × 1.25-mm collimation, pitch of 1.375, 5-mm slice thickness for axial images, and 1.25-mm reconstruction slice thickness, 1.25-mm reconstruction interval, and standard reconstruction algorithm.

From the 179 ROIs, 84 correspond to healthy tissues, 22 to cyst, 32 to carcinoma and the rest to haemangioma instances. Three distinct feature sets were extracted: 32-dimensional GLCM derived features; 7-dimensional FOS features and 9-dimensional temporal features. Cross-validation procedure is used to evaluate the performance of a classifier.

***a.** Full Input Vector*

Full feature set which was composed of all extracted feature subsets fed to the SVM classifier directly without any feature selection, classification Cross Validation (CV) results are listed in table 1.

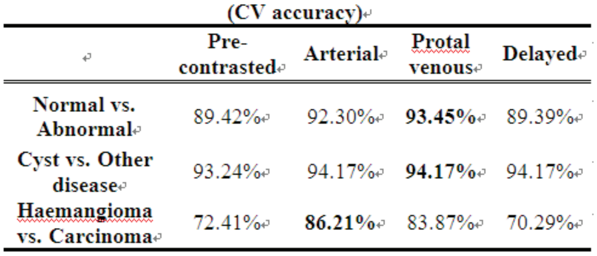
b. Reduced Input Vectors

1) The first step, classifier was trained with the reduced input vectors estimated by applying SFS, SBS, PTA, SFBS and SFFS. To each method, eight features were chosen as a feature vector to train the classifier. Combined these feature subsets together, omitted the repeated features, finally obtained 24 features what were generated the new feature vector. This vector fed to SVM classifier, classification results were shown in table 2.

2) In second step, the important issue in the impact of the algorithm performance is cross-rate P_c and mutation rate P_m. In this study, after a large number of repeated tests, results show that the value of P_c between 0.78 and 0.9, P_m values between 0.005and 0.007 lead to better results. Initial population size was set at 30, the evolution of the largest number of generation was set at 200. As GA is a stochastic method，the optimal features vector includes 10 features estimated via a trial-and-error process, until no further improvement in classification could be obtained. The classification rates are shown in table 3.

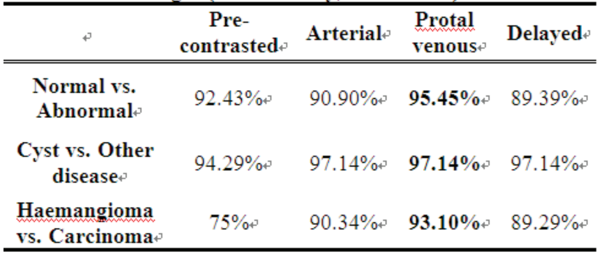
The statistics in the tables reveal that the classification accuracy is 93.45%, 94.17%and 86.21% for normal vs. abnormal, cyst vs. other-disease and carcinoma vs. haemangioma respectively in all the best classification accuracy is achieved using all 48 features. While after first step feature selection, obtained 24 features, the accuracy increased to 95.45%, 97.14%, and 93.10%. Further, for the advantages of GA and SVM, in the second step of the method, the accuracy attained to 98.29%, 98.96% and 96.43% for each sub-problem respectively. From above, we can see that the two-step feature selection method proposed in this paper takes full advantage of the filters in fast computation and the wrappers in high accuracy, which result in lower dimension feature vector and improve the classification performance greatly.

Table1. Full Feature Set Performance in Multiphase Images (CV accuracy)



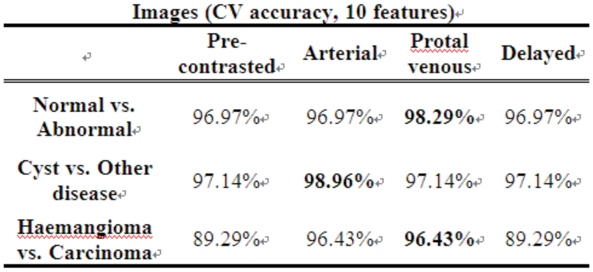
	Pre-contrasted	Arterial	Protal venous	Delayed
Normal vs. Abnormal	89.42%	92.30%	**93.45%**	89.39%
Cyst vs. Other disease	93.24%	94.17%	**94.17%**	94.17%
Haemangioma vs. Carcinoma	72.41%	**86.21%**	83.87%	70.29%

Table2.Reduced Full Feature Set Performance in Multiphase Images (CV accuracy, 24 features)



	Pre-contrasted	Arterial	Protal venous	Delayed
Normal vs. Abnormal	92.43%	90.90%	**95.45%**	89.39%
Cyst vs. Other disease	94.29%	97.14%	**97.14%**	97.14%
Haemangioma vs. Carcinoma	75%	90.34%	**93.10%**	89.29%

Table3.Reduced Feature Subset Performance in Multiphase Images (CV accuracy, 10 features)



	Pre-contrasted	Arterial	Protal venous	Delayed
Normal vs. Abnormal	96.97%	96.97%	**98.29%**	96.97%
Cyst vs. Other disease	97.14%	**98.96%**	97.14%	97.14%
Haemangioma vs. Carcinoma	89.29%	96.43%	**96.43%**	89.29%

IV. Conclusion

The experimental results show that the two-step feature selection module proposed in this paper improved the performance in both speed and accuracy. We also improve in this system in the parameters selection of SVM, further study is needed to optimize kernel parameters and feature subset simultaneously，expecting that one day a liver CAD can be used for clinical.

Acknowledgment

This work is partially supported by the National Fundamental Research Program (973) under Grant No. 2006CB504801 and 2007CB512701 as well as National Natural Science Foundation of China under Grant No. 30770608. The authors would like to thank experts from Shanghai Renji Hospital for their clinical advices.

References

[1] Copyright 2008 © American Cancer Society, Inc.

[2] S.Theodoridis and K.Koutroumbas, Pattern Recognition, third edition: Academic Press, Page 328-335, 2003)

[3] T. Niemeyer, C.Wood, K. Stegbauer, J. Smith, “Comparison of automatic time curve selection methods for breast MR CAD,” SPIE Vol.5370, 2004

[4] W.Siedlechi and J.Sklansky. Constrained Genetic Optimization via Dynamic Reward-Penalty Balancing and Its Use in Pattern Recognition. In proceedings of the International Conference on Genetic Algorithms, Pages 141-150, San Mateo, California, 1989

[5] P.A. Devijer and J. Kittler. Pattern Recognition: a Statistical Approach. Prentice Hall, 1982.

[6] J. Novovicova P. Pudil and J. Kittler. Floating search methods in feature selection. Pattern Recognition, 28(9):1389–1398, 1995.

[7] P. Pudil, J.Novovicova, and J. Kittler. Floating searchmethods in feature selection. Pattern Recognition Letters, 15(11):1119–1125, 1994.

[8] M.J.J.Scott, M.Niranjan, R.W.Prager. “feature subset selection in variable cost domains” CUED/F-INFENG/TR.323 May 1998

[9] P. Langley. “Selection of relevant features in machine learning. In Proceedings of AAAI Fall Symposium on Relevance.” AAAI, September 1994.

[10] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[11] R.H. Duda and P.E. Hart. Pattern classification and scene analysis. Wiley, 1973.

[12] W.Siedlecki and J. Sklansky. A Note on Genetic Algorithm for large-scale Feature Selection. Pattern Recognition Letters, 10(5): 335-347, November 1989

[13] J.Weston, S. and Mu Kherjee. “Feature selection for SVMs” Barnhill BioInformatics.com, Savannah, Georgia, USA.CBCL MIT, Cambridge, Massachusetts, USA. AT&T Research Laboratories, Red Bank, USA. Royal Holloway, University of London, Egham, Surrey, UK.

[14] M. Oren, C. & Papageorgiou. Pedestrian detection using wavelet templates. In *Proc. Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, June 16–20 1997.