USING FEATURE SELECTION WITH SUPPORT VECTOR MACHINE IN GASTRIC HISTOLOGY CLASSIFICATION

C. R. Huang, H. J. Kuo, P. C. Chung and M. Popper

Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

{nckuos, xxray}@neural.ee.ncku.edu.tw, pcchung@ee.ncku.edu.tw, mp-info@ba.sknet.sk

Abstract: This study presented a computer-aided diagnosis system using sequential forward floating selection (SFFS) with support vector machine (SVM) to diagnose gastric histology of Helicobacter pylori (*H. pylori*) from endoscopic images. To achieve the goal, candidate image features associated with clinical symptoms are extracted from endoscopic images. With these candidate features, the SFFS method is applied to select feature subsets, which perform the best classification results under SVM, respect to the different histological features. By using the classifiers obtained from the feature subsets, a new diagnosis system is implemented to provide the physicians with *H. pylori* related histological results during the endoscopy without invasive biopsy

Introduction

Eradication of *Helicobacter pylori* (*H. pylori*) infection has now become central to the management of gastroduodenal diseases [1] [2] [3]. To evaluate the existence of *H. pylori*, the physicians have to obtain several bits of mucosa by gastric biopsy in three parts, antrum, body and cardia, of each patient's stomach. Each specimen is evaluated by pathologists for the *H. pylori*-related gastric histology, of the *H. pylori* density (HPD) [5], the score of *H. pylori*-infection proposed by the updated Sydney system [4]. This histological feature provides clinical significance to monitor the degree of *H. pylori* [5] [6] infection. However, obtaining the gastric histology reguires invasive biopsy, which causes bleeding of mucosa. As a result, only local patches but not the global stomach should be observed. From these patches, obtaining the gastric histology by the updated Sydney system is a complicated and time-consuming procedure burdening to the pathologist. To resolve these disadvantages of histology obtained by using the updated Sydney system, we look forward to recognize the *H. pylori*-infected mucosa from endoscopic images without invasive biopsy.

In general, the gastric mucosa of the *H. pylori*-infection patients will become reddish and un-smooth. Thus, the well-experienced physicians may diagnose the *H. pylori*-infection based on the mucosa variations of color and texture, but such diagnosis will be heavily dependent on personal experience. If the correlations between the endoscopic images and the *H. pylori*-infection from histology can be identified, the diagnosis from the endoscopic images by computers may offer a consistent and objective result. To obtain the correlations, the image features from endoscopic images including color components and texture components [7] [8] are regarded as correlated with the histology according to the physician's experience and therefore, are extracted as the candidate image features from the region of interesting marked by physicians in our approach. Nevertheless, these image features may not present the associations with the histology. Imposing unassociated image features may cause un-convergence in the training stage of a classifier. Even if the classifier converges, the sensitivity and specificity of the classifier will decrease due to the interference of the unassociated image features. To resolve this problem, we use the sequential forward floating selection (SFFS) [10] to choose a subset, which performs the best classification results, of the original image features. In our approach, support vector machine (SVM) [11] [12] [13], which performs significant results dealing with binary classification, is used as the feature selection criterion to decide the existence of *Helicobacter pylori* (*H. pylori*) from the selected subset of images features.

In the remaining part of the manuscript, Section 2 describes how to extract image features from the region of interesting. Feature selection and classification are presented in Section 3. The experimental results are shown in Section 4. Finally, Section 5 provides the conclusions.

Materials and Methods

Patients and study design

A total of 236 patients undergoing diagnostic upper endoscopy due to dyspeptic symptoms were enrolled into this study. Based on a positive result of histology, 130 of the study patients were defined with *H. pylori* infection. To avoid impediments in evaluating *H. pylori* status, patients with the presence of any of the following conditions were excluded: the ingestion of bismuth salts, proton pump inhibitors, or antibiotics in the previous eight weeks; known allergy to penicillin; and previous gastrointestinal surgery. Patients who had histories of anti-*H. pylori* therapy and gastric malignancy were also excluded. After obtaining informed consent from the patients, all patients received gastric biopsies for topographical histology to evaluate the initial status of

ISSN: 1727-1983

H. pylori infection during the endoscopy. Throughout the study, video endoscope (Olympus XQ240, Olympus Corp., Tokyo, Japan) and biopsy forceps (Olympus FB-25N, Olympus Corp.) were applied to obtain three pairs of gastric biopsy sampling from the topographic gastric sites, including antrum (A), body (B), and cardia (C). In each patient, a pair of specimens over the gastric cardia was sampled, under the retroversion of the endoscope, 3 cm within the E-C junction. Next, the other two pairs were obtained from the lower body (on the greater curve) and the antrum (2 cm within the pyloric ring), respectively. According to the regular procedure of endoscopy in National Cheng Kung University Hospital [17], the physicians choose the suspected regions and snap endoscopic images during endoscopy, and then decide if they will perform biopsy in antrum, body and cardia. These endoscopic images were stored at a computer for image analysis as shown in Figure 1. In our approach, the extracted image features will be performed correlation test with histological features obtained by biopsy. In order to obtain the image regions really representing the location of biopsy for analysis, physicians manually selected the regions of interesting (ROI) which they performed biopsy, to be analyzed by our system as the rectangles shown in Figure 1. The parameters describing these rectangle regions were then recorded for the further analysis.

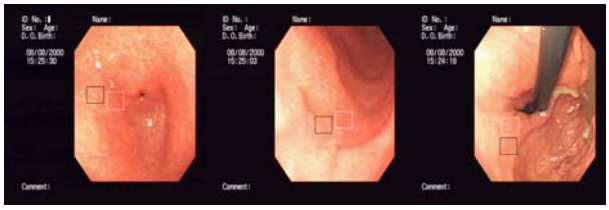

Figure 1: The endoscopic images of stomach represent antrum, body and cardia.

Feature extraction

Throughout this study, the endoscopic images of the biopsy regions from antrum, body, and cardia were analyzed for each patient. Physicians manually selected the regions of interesting (ROI) which they performed biopsy, to be analyzed by our system. The features describing these rectangle regions were then recorded for the further analysis.

Patients who infected *H. pylori* are usually disclosed with reddish and uneven surface of mucosa in clinic. According to the clinical features of mucosa, we developed image features based on two major characteristics (color and texture) observed in endoscopic images. The composition of image features are separate into two major criteria, including aspects related for color and for texture. The image features comprising the color criterion were further defined for the gray-scaled intensity and the components of red, green and blue sub-images. For each sub-image, three statistic features were derived: the maximum, average, minimum, dominant and extension to indicate the maximum value, mean value, minimum value, the value that has the maximum histogram distribution and the number of the histogram bins in the sub image. Accordingly, there are a total of 20 (4×5) RGBI color features.

To represent the smoothness and texture of gastric mucosa surface, we used the algorithm that was proposed by Unser [8] based on sum histograms and difference histograms. The sum and difference associated with the relative displacement (*dx*, *dy*) of pixel (*k*, *l*) are defined as follows:

$$s_{k,l}(d_x,d_y) = I(k,l) + I(k+d_x,l+d_y), \quad (1)$$

and

$$d_{k,l}(d_x,d_y) = I(k,l) - I(k+d_x,l+d_y), \quad (2)$$

where I is the gray level of pixel (k, l). The sum histogram S depends on the displacements dx and dy, and is the histogram of the sums of all pixels. The difference histogram D is the histogram of the differences between pixels that are the specified spatial distance apart. Thus, the sum histogram S and the difference histogram D of the image are defined as follows:

$$S(i) = \#\{(k,l) \in (L_x \times L_y), s_{k,l} = i\}, \quad (3)$$

and

$$D(j) = \#\{(k,l) \in (L_x \times L_y), d_{k,l} = j\}, \quad (4)$$

where # is the number of the elements of the set, L_x and L_y are the horizontal and vertical spatial domains.

To obtain the texture parameters of the endoscopic image, three texture descriptors were generated based on sum histogram and difference histogram as follows:

$$\text{Contrast} = \sum_j j^2 \hat{P}_d(j), \quad (5)$$

$$\text{Entropy} = -\sum_i \hat{P}_s(i)\log(\hat{P}_s(i)) - \sum_j \hat{P}_d(j)\log(\hat{P}_d(j)), \quad (6)$$

and

$$\text{Energy} = \sum_i \hat{P}_s(i)^2 \sum_j \hat{P}_d(j)^2, \quad (7)$$

where $\hat{P}_s(i)$ and $\hat{P}_d(j)$ are the normalized sum histogram and difference histogram, respectively, defined as follows:

$$\hat{P}_s(i) = S(i)/N, i = 2,...,2N_g, \quad (8)$$

and

$$\hat{P}_d(j) = D(j)/N, j = -N_g + 1,..., N_g - 1, \quad (9)$$

where N_g is the total number of the quantified gray levels. In our approach, the displacements of the descriptors in horizontal and in vertical were two-pixel apart. As all sub images are operated by texture descriptors, a total of 120 image parameters (4×2×3×5, four for color, two for direction, three for descriptors, and five for statistic parameters) to indicate the texture of endoscopic images were involved.

Nevertheless, the fixed image resolution cannot be defined so as to cover the majority of the histological results. To further obtain the features of the endoscopic images at arbitrary resolution and direction, we applied discrete wavelet transform (DWT) which is the

ISSN: 1727-1983

fundamental motivation for multi-resolution analysis. The wavelet transform provides a tool for spatial and frequency representation by decomposing the original endoscopic images to the corresponding scale. When decomposition level decreases or increases in the spatial domain, it provides zooming capabilities and local characterization of the image in the frequency domain. In our research, we used biorthogonal CDF 9/7 tap wavelet transform developed by Cohen *et al*. [9] on the endoscopic images. we did 3-level decompositions on endoscopic image from different color models to get ten bands of subimages. From each sub-image (sub-band) of each channel belonging to the color model, we calculated its maximum, average, minimum, dominant and extension of its frequency respectively as image features. In the experiment, we performed discrete wavelet transform on the RGB, HSI and YCbCr model and obtained 450 wavelet image parameters (3×10×5×3, three for color elements, ten for sub-bands, five statistic parameters, and three for color models).

Feature selection and classification

From the previous section, there are 590 candudate image feautures from a endoscopic image. Nevertheless, these image features were only designed statistic values according to physicians' experience in clinic, there is no guarantee that all image features will correlate with *H. pylori*-infection. It would be desirable to obtain the image features, which are correlated with the HPD of the histology. Only the image features correlated with HPD are used for further classification. To extract the significant image features from these image features, which can increase the classification accuracy and decrease the feature number, feature selection is required. In recent researches [14] [15], sequential forward floating selection (SFFS) has been shown efficient in selecting the subset of features [10]. Thus, in our approach, SFFS is used to select the best classification results under SVM.

A. Feature selection

In the sequential forward floating selection procedure, for each forward step, a number of backward steps are also applied to backwardly choose subset which presents better classification performance than the previously evaluated ones at that level. Let Y be the candidate set with size n, of the image features and X be the selected subset of Y, $X \subseteq Y$, with the desired size d. Also let feature selection criterion function for the set X be represented by $J(X)$. In our cases, we consider a higher value of J indicating a better feature subset of classification using SVM. The algorithm of the SFFS [16] can be described as follows:

Input:
$Y = \{y_j \mid j = 1, \ldots, n\}$
Output:
$X_k = \{x_j \mid j = 1, \ldots, k, x_j \in Y\}$, $k = 0, 1, \ldots d$
Initialization:
$X_0 = \Phi$, and $k = 0$
Step 1 (Inclusion)
$x^+ = arg\ max\ J(X_k + x), x \in Y - X_k$
$X_{k+1} = X_k + x^+$; $k = k + 1$
Step 2 (Conditional exclusion)
$x^- = arg\ max\ J(X_k - x), x \in X_k$
if $J(X_k - \{x^-\}) > J(X_{k-1})$ then
$X_{k-1} = X_k - x^-$; $k = k - 1$
Go to Step 2
Else
Go to step 1
Step 3 (Termination)
When $k = d$, stop the procedure

After SFFS, X_d is the subset of features that performs the best classification results according to the criterion function using SVM.

B. Feature classification

In our approach, support vector machine (SVM) [11] [12] [13], which minimizes the upper bound of the generalization error, is used to dealing with the binary classification problem. Let $\boldsymbol{x}_i \in R^n$, i = 1, …,N be the training vectors in two classes and $y_i \in R^N$ be the observation results such that $y_i \in \{1, -1\}$. Each feature vector $\boldsymbol{x}_i$ is transformed to a high dimensional feature space according a nonlinear mapping $\Phi(\boldsymbol{x})$, $\boldsymbol{z} = \Phi(\boldsymbol{x})$ and then SVM searches the separating hyperplane $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z} + b$ in the feature space with the largest margin. The $\boldsymbol{w}$ for this optimal hyperplane can be written as $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{z}_i$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ can be found by solving the following quadratic programming (QP) problem:

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Q}\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha} \tag{10}$$

subject to

$$C \geqq \alpha_i \geqq \mathbf{0}, \text{ and } \boldsymbol{\alpha}^T \boldsymbol{Y} = 0, \tag{11}$$

, where $C > 0$ is the upper bound, $\boldsymbol{e}$ is the vector which contains all ones, $\boldsymbol{Q}$ is a symmetric $N \times N$ matrix with $Q_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j)$ is the kernel function. In our approach, we use radial-basis function (RBF)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / \sigma} \tag{12}$$

as the kernel functions to classify our data.
During the testing, for a test vector $\boldsymbol{x} \in R^n$, we use the decision function shown as follows:

$$sign\left(\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right) \tag{13}$$

to obtain the class label.

Results

In the experiments, the HPD from histology is regarded as the ground truth of the presence of *H.*

ISSN: 1727-1983

pylori-infection. We divided the data into two randomly selected subsets for training and testing and performed 10 fold cross validation of our data for avoiding the over-training. To select a subset of image features from the training set, SFFS was applied first. The feature selection criterion function is the classification accuracy of SVM using the software LIBSVM [18], which provides the support vector classification with the RBF kernel. The procedures of SFFS on selecting our image features for antrum, body and cardia are shown in Figure 2, Figure 3 and Figure 4, respectively. As a comparison, we also applied use sequential forward selection (SFS) to select the image features. In general, as more image features are added, there is a relatively smooth rise in the recognition rate until the peaks reach, and then eventually falls with dimension of features. In these figures, due to conditional exclusion of SFFS, the number of features in the subset may decrease. Thus, in some feature subset, there are more than one classification accuracies. In contrast, the number of features in the subset using SFS is continuously increased. From the experiments, although both SFFS and SFS can obviously decrease the number of elements of the input vector for classification, the feature subsets selected by SFFS perform better accuracy in most cases.

Then we used the selected subset, which performs the best classification accuracy as new feature set and applied SVM with polynomial and RBF kernels for training and testing. For comparison, we also used total 590 image features as the elements of the input vector to SVM [18]. In all experiments, the constant C controlling the tradeoff between training error and model complexity is 1. By evaluating the testing set, classification accuracy (Acc), sensitivity (SE), and specificity (SP) were computed as follows:

$$\text{Acc} = \frac{TP + TN}{N}, \tag{14}$$

$$\text{SE} = \frac{TP}{N_P}, \tag{15}$$

and

$$\text{SP} = \frac{TN}{N_N}, \tag{16}$$

where N is the total number of the patients, TP is the number of the positive patients, N_P is the number of the positive patients who have infected *H. pylori* according to the histology, TN is the number of the negative patients, N_N is the number of the patients who does not infect *H. pylori* according to the histology. The average accuracies, sensitivities, specificities and the number of selected subset features (# of SFs) used in the input vectors for classifying HPD in antrum, body, and cardia after 1000 times of evaluations using SVM with RBF kernel and SFFS with SVM (SFFS+RBF) are shown in Table 1, Table 2 and Table 3, respectively.

From Table 1 to Table 3, we can see that using SFFS to select the subset of images features will obviously decrease the number of elements of the input vector for classification. Such results will also decrease the training and testing time. Moreover, the average accuracies using proposed SFFS with SVM, which uses much less number of the input feature vector, performs better classification results than the results of SVM with full candidate image feature vector. Also shown in these tables, the proposed SFFS+SVM performs good sensitivities to detect the presence of *H. pylori*-infection. However, the specificities are much lower than the sensitivities, which is will misclassify the normal cases that do not infect *H. pylori* to abnormal cases.

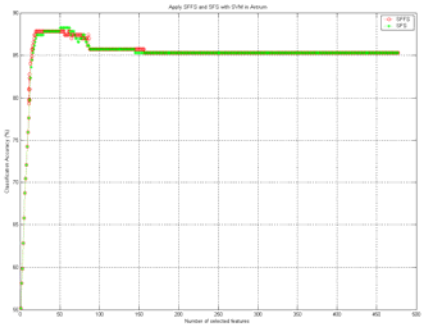

Figure 2: Classification accuracy of SFFS and SFS with SVM respective to the image features of antrum.

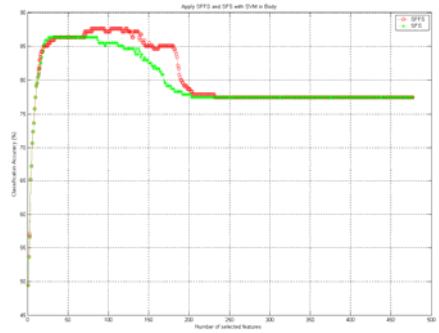

Figure 3: Classification accuracy of SFFS and SFS with SVM respective to the image features of body.

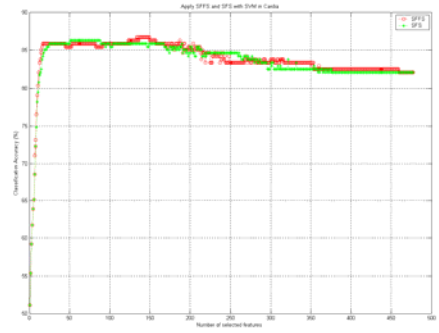

Figure 4: Classification accuracy of SFFS and SFS with SVM respective to the image features of cardia.

ISSN: 1727-1983

Table 1: Testing results of SVM and SFFS with SVM in the antrum of the stomach.

	Acc(%)	SE(%)	SP(%)	# of SFs
RBF	86.3	98.5	70.8	450
SFFS+RBF	87.3	99.3	73.0	63

Table 2: Testing results of SVM and SFFS with SVM in the body of the stomach.

	Acc(%)	SE(%)	SP(%)	# of SFs
RBF	86.4	98.7	71.5	450
SFFS+RBF	87.4	99.8	73.6	123

Table 3: Testing results of SVM and SFFS with SVM in the cardia of the stomach.

	Acc(%)	SE(%)	SP(%)	# of SFs
RBF	86.0	98.1	70.3	450
SFFS+RBF	86.2	99.2	71.9	149

As we obtained the feature subsets which can perform better classification accuracy than the original candidate image features, we also computed the controbution of different type of image features in different gastric parts as shown in Table 4. The image features computed from HSI model with discrete wavelet transform perform the most significant evidences for *H. pylori* diagnosis.

Table 4: Percentage of selected image features using color and texture, RGB model with discrete wavelet transform (RGB with DWT), HSI model with discrete wavelet transform (HSI with DWT) and YCbCr model with discrete wavelet transform (YCbCr with DWT) in antrum, body and cardia parts.

Features	Antrum	Body	Cardia
Color and Texture (%)	4.76	0.00	2.02
RGB with DWT (%)	30.16	29.27	22.82
HSI with DWT (%)	57.14	50.40	61.07
YCbCr with DWT (%)	7.94	20.33	14.09

Conclusions

In this study, we presented a computer-aided diagnosis system by SFFS with SVM to analyze the endoscopic images and report the presence of *H. pylori*-infection of the patients. As shown in the experiments, SFFS with SVM is disclosed with significant effectiveness in predicting the presence or absence of *H. pylori* infection. Our system can provide the physician an assistant opinion in the diagnosis of the *H. pylori*-related histology without invasive biopsy. Moreover, taking images around the global stomach will not injure the mucosa of the patient, and consequently the system also indirectly improves the local biopsy into global evaluation of the whole stomach. Considering the effects on the endoscopic images such as shadows of endoscope, gastric juices and gastric bleeding, the selection of regions of interesting is performed manually. Automatic ROI selection remains as the future work.

Acknowledgment

The authors would like to thank Dr. Bor-Shyang Sheu, Dr. Hsiao-Bai Yang and Miss Hunt-Wei Wu for the assistance. The work was supported by the National Health Research Institute, Taiwan, under Grant NHRI-EX90-8941SC, and the National Science Council, Taiwan, under Grant NSC 91-2622-E-006-079.

References

[1] J. PARSONNET, G. F. FRIEDMAN, D. P. VANDERSTEEN, Y. CHANG, J. H. VOGELMAN, N. ORENTREICH, and R. K. SIBLEY, "*H. pylori* infection and the risk of gastric carcinoma," *N. Engl. J. Med.*, vol. 325, pp. 1127-1131, 1991.

[2] J. H. WALSH, and W. L. PETERSON, "The treatment of *Helicobacter pylori* infection in the management of peptic ulcer disease," *N. Engl. J. Med.*, vol. 333, pp. 984-991, 1995.

[3] E. M. EL-OMAR, I. D. PENMAN, J. E. ARDILL, R. S. CHITTAJALLU, C. HOWIE, and K. E. MCCOLL, "*Helicobacter pylori* infection and abnormalities of acid secretion in patients with duodenal ulcer disease," *Gastroenterology*, vol. 109, pp. 681-691, 1995.

[4] M. F. DIXON, R. M. GENTA, J. H. YARDLEY, P. and CORREA, "Classification and grading of gastritis. The updated Sydney system," *International Workshop on the Histopathology of Gastritis*, Houston, 1994.

[5] B. S. SHEU, S. B. YANG, I. J. SU, C. H. CHI, S. C. SHIESH, and X. Z. LIN, "Bacterial density of *H. pylori* predicts the success of triple therapy in bleeding duodenal ulcer," *Gastrointestinal Endoscopy*, vol. 44, pp. 683-688, 1996.

[6] B. S. SHEU, J. J. WU, H. B. YANG, A. H. HUANG, and X. Z. LIN, "One week proton pump inhibitor-based triple therapy is effective in eradicating residual *H. pylori* after failed dual therapy," *J. Formos. Med. Assoc.*, vol. 97, pp. 266-270, 1998.

[7] R. M. HARALICK, K. SHANMUGAM, and I. DINSTEIN, "Textural features for image classification," *IEEE Trans. on System. Man. Cybernetics*, vol. 3, no. 6, pp. 610-621, 1973.

[8] M. UNSER, "Sum and difference histograms for texture classification," *IEEE Trans. Patt. Ana. Mach. Intell.*, vol. 8, no.1 , pp. 118-125, 1986.

[9] A. COHN, I. DAUBECHIES and J.C.FEAUVEAU, Biorthogonal Bases of Compactly supported wavelets, *Comm. Pure and Applied Math*, 1992

[10] P. PUDIL, J. NOVOVIČOVÁ, and J. KITTLER,

"Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1199-1125, 1994.

[11] V. VAPNIK, The nature of statistical learning theory, Berlin, Germany: Springer-Verlag, 1995.

[12] CORTES and V. VAPNIK, "Support-vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[13] C. J. C. BURGES, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 955–974, 1998.

[14] A. JAIN and D. ZONGKER, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Patt. Ana. Mach. Intell.*, vol. 19, no. 2, 1997.

[15] M. KUDO and J. SKLANSKY, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25-41, 2000.

[16] P. PUDIL, F. J. FERRI, J. NOVOVICOV'A, and J. KITTLER, "Floating search methods for feature selection with nonmonotonic criterion functions," *Proceedings of International Conference on Pattern Recognition*, pp. 279-283, 1994.

[17] S. B. YANG, B. S. SHEU, I. J. SU, C. H. CHIEN, and X. Z. LIN, "Clinical application of gastric histology to monitor treatment of dual therapy in *H. pylori* eradication," *Dig. Dis. Sci.*, vol. 42, pp. 1835-1840, 1997.

[18] C.-C CHANG, and C.-J. LIN, LIBSVM: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2003.

ISSN: 1727-1983