AN AUTOMATIC CLASSIFICATION SYSTEM APPLIED IN MEDICAL IMAGES

Bo QIU, Chang Sheng XU, Qi TIAN

Institute for Infocomm (I2R), Singapore, 119613
{qiubo,xucs, tian}@i2r.a-star.edu.sg

ABSTRACT

In this paper, a multi-class classification system is developed for medical images. We have mainly explored ways to use different image features, and compared two classifiers: Principle Component Analysis (PCA) and Supporting Vector Machines (SVM) with RBF (radial basis functions) kernels. Experimental results showed that SVM with a combination of the middle-level blob feature and low-level features (down-scaled images and their texture maps) achieved the highest recognition accuracy. Using the 9000 given training images from ImageCLEF05, our proposed method has achieved a recognition rate of 88.9% in a simulation experiment. And according to the evaluation result from the ImageCLEF05 organizer, our method has achieved a recognition rate of 82% over its 1000 testing images.

1. INTRODUCTION

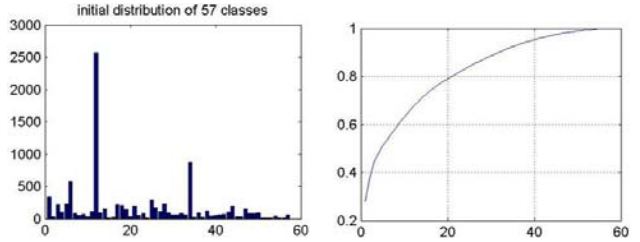
With the fast development of modern medical devices, more and more medical images are generated, so that the demand becomes more and more urgent for automatically indexing, comparing, analyzing and annotating the huge volume of medical images. Medical images are a kind of medical evidence to patients and doctors. To interpret those medical evidences, generally doctors will use specialist vocabulary and natural language phrases, and relate them to some specific cases. It is difficult for some unskilled doctors but automatic annotation of medical images will do much help to them.

For automatic annotation, which is a kind of automatic machine-based reasoning based on the evidence gathered, additional interpretive semantics must be attached to the image data. About this some methods have been explored in special domains, like the diagnosis of breast cancer [1]. But until now in a wider domain, there is no popular method for automatic annotation owing to the variety of medical images and the lack of relevant domain knowledge. So in this paper we simplify the problem into a multi-class classification problem, which means that the classification labels assigned to the classes are regarded as a simple annotation.

According to [2], classification methods include parametric and nonparametric. With given training data, in this paper only parametric methods are considered, which includes Bayesian estimation (Maximum-Likelihood, Hidden Markov models, Expectation-Maximization, Fisher Linear Discriminant, Multiple Discriminant Analysis, etc.), Linear Discriminant functions (Perceptron Criterion Function, Relaxation Procedures, Minimum Squared-Error Procedures, PCA, SVM, Ho-Kashyap Procedures, etc.), Multi-layer Neural Networks, Stochastic methods (Simulated Annealing, Boltzmann learning, Evolutionary methods, etc.).

The methods above have been applied successfully in many fields [2]. But until now the problem of medical images classification is a new and great challenge, because when compared with other classification problems, there are some particular difficulties in medical images:

- *Great unbalance between classes*
  Figure 1 shows the size of each class in our database (see experiment part). It can be found that, class 6 has more than 500 samples, class 12 has more than 2,500 samples, class 34 has near 1,000 samples, while all the others are much less — the minimal class has only 9 samples. 20 largest classes occupy near 80% of the whole dataset. This unbalance makes many common classification methods unavailable.



(a) Sizes vs. classes (b) Size percentages vs. classes

Figure 1. Great unbalance between classes

- *Visual similarities between some classes (See Figure 2)*
  Unlike the other image databases, for medical images, sometimes even skilled experts cannot find the differences between some classes visually. They may need to compare the images from different sources and refer to other medical examinations like blood.

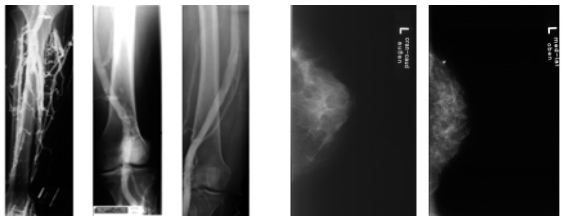


Figure 2. Visual similarities between some classes

- *Variety in one class and difficulty to define discriminative visual features (See Figure 3)*
  Too many modalities vary in one class. To find a general visual feature for one class is often very difficult. In many cases, medical similarities are far away from visual similarities.

Figure 3. Variety in one class

To face the difficulties mentioned above, based on our former work [7], PCA and SVM are chosen as classifiers in this paper. And different features from low-level to middle-level are considered. Our contributions are:

- Construct a multi-class classification system for medical images;
- Find the most efficient features for classification by designed simulation experiments (some training data are used to simulate testing data).

2. FEATURE SETS

Feature extraction is a basic problem in image processing field. After reviewing 56 CBIR (content-based image retrieval) systems, in [3] a summary of low-level features are listed in 3 main categories: color, texture, and shape, plus a single features: layout. In [4][5] there are some similar categories of features.

The feature 'layout' is the absolute or relative spatial position of the color. It may include low-resolution-pixel-map (LRPM), which is used in our method. LRPM is a down-scaled image of an initial one.

In our system texture maps are calculated on both initial images and filtered images. Filtered images are generated from initial ones by filters like Gaussian, to minimize the influence of noises. Moreover, texture histogram is calculated on these texture maps. Figure 4 shows an example of textures and LRPM.

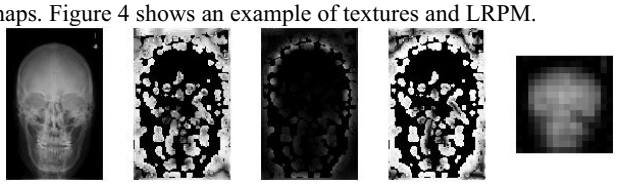


Figure 4. An initial image and its feature maps

Besides the low-level features like LRPM and texture, middle-level regional features such as Blob are also considered [6]. Blob has been applied successfully in medical image retrieval in our past work [7]. Its parameters include: color, texture, area, length of long and short axes, rotation angle, Fourier decomposition parameters, etc.

Because there are so many features available, feature selection becomes a key problem. The 'best' features should be the most distinguishing features, and invariant to irrelevant transformations of the input. Facing all kinds of features, it is extremely difficult to find out which are the best ones theoretically. The practical way is to select suitable features by simulation experiments, where the best features can lead to the best classification results.

3. CLASSIFICATION METHODS

3.1. Classifiers: SVM and PCA

SVM is widely used for statistical learning, classifiers and regression models design [8]. Primarily SVM tackles the binary classification problem [9]. According to [10], SVM for multiple-classes classification is still under development, and generally there are two types of approaches. One type has been to incorporate multiple class labels directly into the quadratic solving algorithm. Another more popular type is to combine several binary classifiers. We use SVM^{Torch}, which belongs to the latter.

Kernel selection is a crucial issue for SVM. Different kernels will accommodate different nonlinear mappings and the performance of the resulting SVM will often hinge on the appropriate choice of the kernel [11]. There are 4 kernels in SVM^{Torch}: linear, polynomial, radial basis function (RBF), sigmoid tanh. In our method RBF is chosen:

$$K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}. \quad (1)$$

Besides the standard variance σ, another parameter is the trade-off between training error and the margin C.

To compare different methods' effects, PCA is also applied in our experiments. A conventional PCA process starts from its generating matrix's construction. Given a vector dataset (training dataset including *n* images):

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T, \quad (2)$$

in which $\mathbf{x_i}$ (i = 1,2,…,n) can be regarded as an image vector with the length *p* (represents the number of features), PCA generating matrix can be constructed as follows:

$$\mathbf{C}_1 = \mathbf{X}\mathbf{X}^T \text{ or } \mathbf{C}_2 = \mathbf{X}^T\mathbf{X}, \quad (3)$$

where the size of $\mathbf{C_1}$ is nxn, and the size of $\mathbf{C_2}$ is *pxp*. If *n* is lowered down with $\mathbf{C_1}$ after PCA process, it will result the generation of PCA templates to represent the whole dataset; otherwise when $\mathbf{C_2}$ is applied and *p* is lowered down, PCA will reduce the feature vectors' dimension. What we used in this paper is PCA template. Thus the distance measurement is between an unclassified image vector and the templates of different classes.

3.2. System Structure

Figure 5 shows the flowchart of our system, which includes two stages: training and testing.

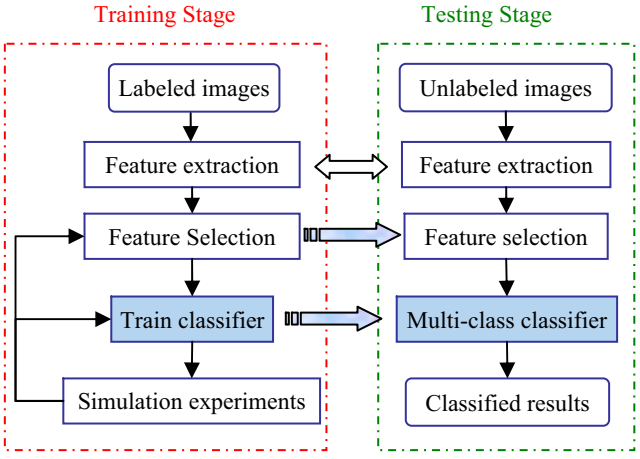


Figure 5. Flowchart of our system

In training stage, the purpose is to do feature selection and train a multi-class classifier. The results from the simulation experiment are used to select the best features and adjust the

parameters of the classifier. In testing stage, the trained multi-class classifier is applied to classify the unlabeled testing images.

In this structure, classifier is not limited to SVM or PCA. This will enhance the system's adaptability. The 'simulation experiment' means that in the experiment the system divides the labeled training images into 2 parts, inside which, one is used as training dataset; the other is used as testing dataset. Obviously this testing dataset is ground-truth-known. By evaluating the results of simulated experiments, we can choose the best features, suitable classifier, and their best parameters. After all, different divisions of the labeled training images will cause different results. Thus the simulation experiments should proceed under various divisions.

4. EXPERIMENTS AND RESULT ANALYSIS

As a benchmark project, ImageCLEF is more and more well-known with its open data platform [12]. And the database used in our experiments is coming from ImageCLEF 2005. In its radiograph database, there are totally 9,000 training images belonging to 57 classes as shown in Figure 6; and 1,000 unlabeled radiographs are given as testing dataset. The task is clear: using training dataset to construct a multi-class classifier for automatic labeling the 1,000 testing images. Evaluation of the system will base on 'recognition rate', which means the percent of how many images are correctly classified. In our case it is equal to average accuracy (AA):

$$AA = \frac{number\_of\_all\_right\_classified\_images}{size\_of\_whole\_testing\_dataset} \quad (4)$$

Figure 6. 57 given classes in ImageCLEF 2005

Our solution starts from simulation experiments to make the selection of features, methods, and parameters based on AA, and goes to the true experiments to classify given testing images.

4.1. Simulation experiment

In each of 57 training sets, 80% images are taken as training data and the left 20% are regarded as testing data. Of course different division of the training sets will cause different results. Here only the division with the highest AA is shown.

- **Simulation experiment for PCA**

First of all, owing to our past work [7], PCA method is chosen as the classifier. PCA combined with Blob features made a good result in image retrieval, but when applied in this task, the AA is rather low (only 50.26%, see Figure 7).

4 statistic column drawings are in Figure 7, Figure 9 and Figure 10: horizontal axis is marked from 1 to 57 (labels of classes), vertical axes are (in order) correctness of each class, number of wrong classified images of each class, number of classified images of each class, number of true images of each class.

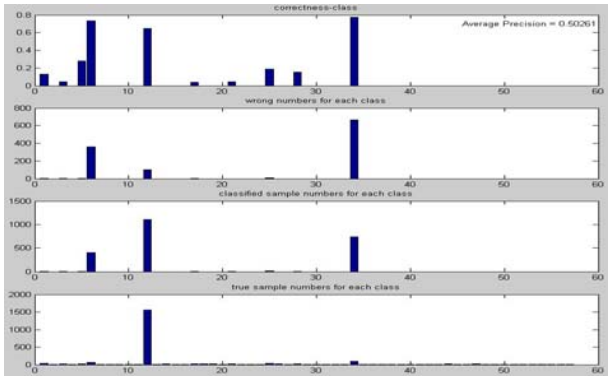


Figure 7. Classification result: PCA + Blob

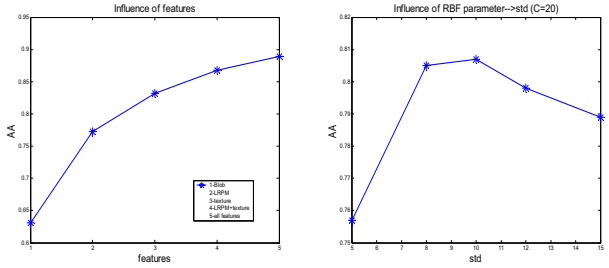
PCA plus other features such as texture is neither successful. This is owing to the over-fitting problem caused by the unbalance of 57 classes, which means that some 'big' classes have too much training data while the others have little, so that in the testing stage too many images are wrongly classified to those 'big' classes. As we can see in Figure 7, most of the classes haven't been recognized and most of the wrongly classified images are labeled as class 6 and 34.

- **Simulation experiment for SVM**

Soft margin SVM is tested with different features:

1) SVM + Blob
2) SVM + LRPM
3) SVM + texture
4) SVM + LRPM + texture
5) SVM + Blob + LRPM + texture

In all of them, 'SVM + Blob + LRPM + texture' reaches the highest AA (88.9%, see Figure 8(a) and Figure 9), and the best variance σ (SVM parameter) is 0.20 (see Figure 8(b)). Here the texture means the down-scaled texture maps calculated based on filtered images.



(a) AA vs. features (b) AA vs. σ

Figure 8. Features and parameter selection

- **Comparisons**

Comparing from the simulation experiments, in the methods part we can see that SVM is better than PCA because SVM can reach higher AA and 'recognize' more classes (for PCA many classes' accuracies are zero); in the features part, we can find the

most effective features are the combination of low-level features (LRPM + texture) and middle-level feature (Blob).

With the help of Figure 8(b), the best σ is defined to 0.20.

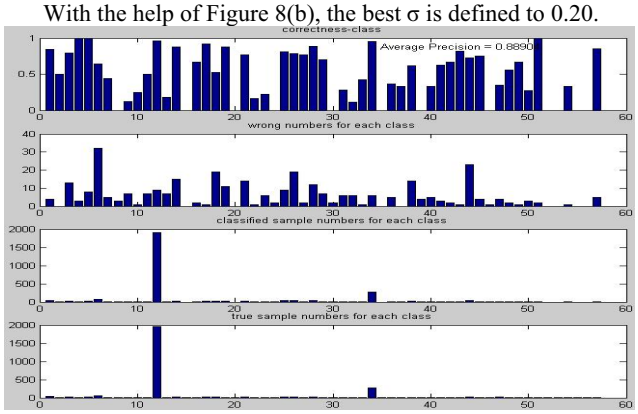


Figure 9. Classification result: SVM + LRPM + texture + Blob

4.2. True experiment

Derived from simulation experiments, the 'SVM + Blob + LRPM + texture' method is chosen at last, which is proved to be the best combination of method and features.

With the given 1,000 testing data, the final AA is 82% (see Figure 10), which means in total there are 820 testing images classified correctly. Meanwhile, there are 23 classes whose AAs are lower than 50%, relevant to 129 images; and there are 11 classes whose AAs are higher than 90%, relevant to 515 images. This shows the influence of unbalance.

Figure 11 is the precision-recall graph of the 57 classes. Each point represents a class. 'G' means 'good region', and the more classes fall in it, the better the result is. As for our results, there are 53% classes falling in G.

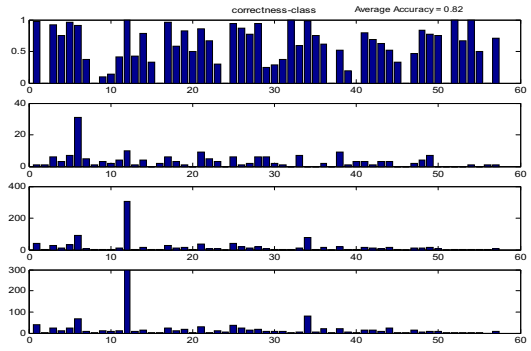


Figure 10. Last result of classification

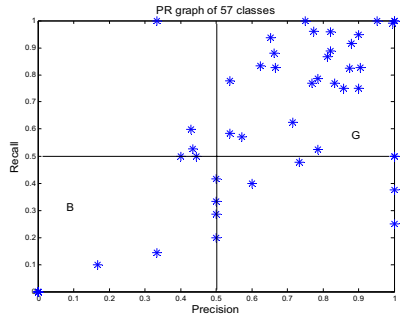


Figure 11. PR graph (G: good region; B: bad region)

5. CONCLUSIONS AND FUTURE WORK

From the experiment results, we can see that SVM behaves better than PCA in constructing a multi-class classifier for this medical image database, owing to the over-fitting problem caused by unbalance of classes. 'SVM + Blob + LRPM + texture' method reaches an AA of 82%, with the kernel RBF. SVM variance influences the result greatly and its best value is 0.20. PR graph is introduced to judge the effects of classification algorithms.

Future work will focus on solving the unbalance problem, and testing new classifiers like neural networks.

6. REFERENCES

[1] B.Hu, S.Dasmahapatra, P.Lewis, and N.Shadbolt, Ontology-based Medical Image Annotation with Description Logics, *Proceedings of The 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 77-82, Sacramento.

[2] Richard O.Duda, Peter E.Hart, David G.Stork, Pattern Classification, A Wiley-Interscience Publication, 2nd ed., ISBN 0-471-05669-3, 2000.

[3] Remco C. Veltkamp, Mirela Tanase, Content-Based Image Retrieval Systems: A Survey, Technical Report UU-CS-2000-34, Oct. 2000, http://give-lab.cs.uu.nl/cbirsurvey/.

[4] T. Lehmann et al, Automatic Categorization of Medical Images for Content-based Retrieval and Data Mining, *Computerized Medical Imaging and Graphics*, vol. 29, pp. 143-155, 2005.

[5] Björn Johansson, A Survey on: Contents Based Search in Image Databases, LiTH-ISY-R-2215, Technical Reports from the Computer Vision Laboratory, Dept. of Electrical Engineering, Linköping University, Sweden, Feb., 2000.

[6] C. Carson et al, Blobworld: A system for region-based image indexingand retrieval, *Proceeding of Third International Conference Visual Information Systems*, 1999.

[7] Wei Xiong, Bo Qiu, Qi Tian, Henning Müller, Changsheng Xu, A novel content-based medical image retrieval method based on query topic dependent image features (QTDIF), *SPIE Medical Imaging,* San Diego, CA, USA, 2005.

[8] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining Knowledge Discovery*, 2:121—167, 1998.

[9] K.-S. Goh, E. Chang, K.-T. Cheng, Support vector machine pairwise classifiers with error reduction for image classification, *Proceedings of ACM MIR*, The Association for Computing Machinery, Ottawa, Canada, pp. 32-37, 2001.

[10] R.Collobert, S.Bengio, SVMTorch: support vector machines for large-scale regression problems, *The Journal of Machine Learning Research,* Vol.1, pp.143-160, 2001.

[11] Tony Jebara, Multi-task feature and kernel selection for SVMs, *Proceedings of the twenty-first international conference on Machine learning ICML '04*, July 2004.

[12] P.Clough, H.Müller, M.Sanderson, The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004, *CLEF Proceedings - Springer Lecture Notes in Computer Science*, LNCS 3491, pp. 597-613, 2005.