Automated X-Ray Image Annotation

Single versus Ensemble of Support Vector Machines*

Devrim Unay[1], Octavian Soldea[2], Sureyya Ozogur-Akyuz[3], Mujdat Cetin[2], and Aytul Ercil[2]

[1] Electrical and Electronics Engineering, Bahcesehir University, Istanbul, Turkey
[2] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey
[3] Mathematics and Computer Science, Bahcesehir University, Istanbul, Turkey
devrim.unay@bahcesehir.edu.tr, octavian@sabanciuniv.edu, sureyya.akyuz@bahcesehir.edu.tr,{mcetin, aytulercil}@sabanciuniv.edu

Abstract. Advances in the medical imaging technology has lead to an exponential growth in the number of digital images that needs to be acquired, analyzed, classified, stored and retrieved in medical centers. As a result, medical image classification and retrieval has recently gained high interest in the scientific community. Despite several attempts, such as the yearly-held ImageCLEF Medical Image Annotation Challenge, the proposed solutions are still far from being sufficiently accurate for real-life implementations.

In this paper we summarize the technical details of our experiments for the ImageCLEF 2009 medical image annotation challenge. We use a direct and two ensemble classification schemes that employ local binary patterns as image descriptors. The direct scheme employs a single SVM to automatically annotate X-ray images. The two proposed ensemble schemes divide the classification task into sub-problems. The first ensemble scheme exploits ensemble SVMs trained on IRMA sub-codes. The second learns from subgroups of data defined by frequency of classes. Our experiments show that ensemble annotation by training individual SVMs over each IRMA sub-code dominates its rivals in annotation accuracy with increased process time relative to the direct scheme.

1 Introduction

Digital medical images, such as standard radiographs (X-Ray) and computed tomography (CT) images, represent a large part of the data that need to be stored, archived, retrieved, and shared among medical centers. Manual labeling of this data is not only time consuming, but also error-prone due to inter/intra-observer variations. In order to realize an accurate classification of digital medical images one needs to develop automatic tools that allow high performance image annotation, i.e. a given image is automatically labeled with a text or a code without any user interaction.

* This work was supported in part by the Marie Curie Programme of the European Commission under FP6 IRonDB project MTK-CT-2006-047217.

Several attempts in the field of medical images have been performed in the past, such as the WebMRIS system [1] for cervical spinal X-Ray images, and the ASSERT system [2] for CT images of the lung. While these efforts consider retrieving a specific body part only, other initiatives have been taken in order to retrieve multiple body parts.

The yearly held ImageCLEF Medical Image Annotation challenge, run as part of the Cross-Language Evaluation Forum (CLEF) campaign, aims in automatic classification of an X-Ray image archive containing more than 12,000 images randomly taken from the medical routine. The dataset contains images of different body parts of people from different ages, of different genders, under varying viewing angles and with or without pathologies.

A potent classification system requires the image data to be translated into a more compact and more manageable representation containing descriptive features. Several feature representations have been investigated in the past for such a classification task. Among others, image features, such as average value over the complete image or its sub-regions [3] and color histograms [4], have been investigated. Recently in [5], texture features like local binary patterns (LBP) [6] have been shown to outperform other types of low-level image features in classification of X-Ray images. Subsequently in [7], it has been shown that retaining only the relevant local binary pattern features achieves comparable classification accuracies with smaller feature sets, thus leading to reduced processing time and storage space requirements.

A less investigated path is to exploit from hierarchical organization of medical data, such as the ImageCLEF data labeled by the IRMA coding system, using ensemble classifiers. Accordingly, in this paper we explore the annotation performance of two ensemble classification schemes based on IRMA sub-codes and frequency of classes, and compare them to the well-known single-classifier scheme over the ImageCLEF-2009 Medical Annotation dataset.

The paper is organized as follows. Section 2 presents our feature extraction and classification steps in detail. Then, in Section 3 we introduce the image database and the experimental evaluation performed. And finally, Sections 4 and 5, present corresponding results and our conclusions, respectively.

2 Method

2.1 Feature Extraction

We extract *spatially enhanced local binary patterns* as features from each image in the database. LBP [6] is a gray-scale invariant local texture descriptor with low computational complexity. The LBP operator labels image pixels by thresholding a neighborhood of each pixel with the center value and considering the results as a binary number. Formally, given a pixel at (x_c,y_c), the resulting LBP code can be expressed as:

$$LBP_{P,R}(x_c, y_c) = \sum_{n=0}^{P-1} s(i_n - i_c)2^n \tag{1}$$

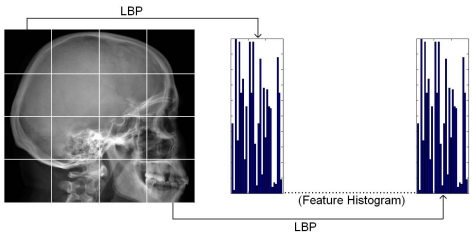

Fig. 1. The image is divided into 4x4 non-overlapping sub-regions from which LBP histograms are extracted and concatenated into a single, spatially enhanced histogram.

where n runs over the P neighbors of the central pixel, i_c and i_n are the gray-level values of the central and the neighboring pixels, and $s(x)$ is 1 if x $\geq$ 0 and 0 otherwise.

Eventually, a histogram of the labeled image $f_l(x, y)$ can be defined as

$$H_i = \sum_{x,y} I(f_l(x, y) = i), \qquad i = 0, \ldots, L-1 \tag{2}$$

where L is the number of different labels produced by the LBP operator, and $I(A)$ is 1 if A is true and 0 otherwise.

The derived LBP histogram contains information about the distribution of local micro-patterns, such as edges and flat areas, over the image. Because not all LBP codes are informative [6], we use uniform version of LBP and reduce the number of informative codes from 256 to 59 (58 informative bins + one bin for noisy patterns). In order to obtain a more local description, we divide images into 4x4 non-overlapping sub-regions and concatenate the LBP histograms extracted from each region into a single, spatially enhanced feature histogram, as in [5] (Figure 1).

Finally, we obtain a total of 944 features per image, and each feature is linearly scaled to [-1,+1] range before presented to the classifier.

2.2 Image Annotation

In this work we use a support vector machine (SVM) based learning framework to automatically annotate the images. SVM [8] is a popular machine learning algorithm that provides good results for general classification tasks in the computer vision and medical domains: e.g. nine of the ten best models in ImageCLEFmed 2006 competition were based on SVM [9]. In a nutshell, SVM maps data to a higher-dimensional space using kernel functions and performs linear discrimination in that space by simultaneously minimizing classification error and maximizing geometric margin between classes.

Among all available kernel functions for data mapping in SVM, Gaussian radial basis function is the most popular choice, and therefore it is used here.

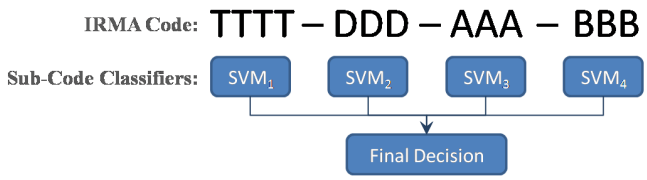

Fig. 2. Illustration of ensemble classification based on IRMA sub-codes. A separate SVM is trained for each sub-code, and final decision is formed by concatenating predictions of each SVM.

In this work we used LibSVM[1] library (version 2.89) for SVM and empirically found its optimum parameters on the dataset.

Direct Annotation Scheme (D). In this scheme, we classify images by using a single SVM with one versus all multi-class model.

To the contrary, ensemble schemes break down the annotation task to sub-problems by dividing the data into subgroups based on 1)IRMA sub-codes, and 2)frequency of classes.

Ensemble Annotation by IRMA sub-codes (E-1). In the IRMA coding system, images are categorized in a hierarchical manner based on four sub-codes describing image modality, image orientation, body region examined, and biological system investigated. Accordingly, in this scheme we train a separate SVM for each sub-code and merge their predictions to form the final decision, as illustrated in Figure 2.

Ensemble Annotation by frequency of classes (E-2). On the contrary, this ensemble scheme successively divides the data into sub-groups based on frequency of classes and trains a separate SVM on each sub-group (Figure 3). Let $L_1, L_2, \ldots, L_n$ be the set of classes in the training set and $m \in N$ be a positive integer parameter. Without loss of generality, assume $L_1, L_2, \ldots, L_n$ are sorted in their decreasing cardinality values. We divide the training set in a sequence clusters $C_1, C_2, \ldots, C_k$, such that $C_1 = \{L_1, L_2, \ldots, L_m, U_1\}$, $C_2 = \{L_{m+1}, L_{m+2}, \ldots, L_{2m}, U_2\}$, where $U_1 = \bigcup_{i=m+1}^{n} L_i, U_2 = \bigcup_{i=2m}^{n} L_i$, and so on, see Figure 3. For each C_i we train a SVM. Let S_i be the SVM trained on C_i. When classifying, we begin from S_1. If S_1 suggests one of the $L_1, L_2, \ldots, L_m$ labels, then we consider this result a valid classification. If the result is U_1, then we proceed further to S_2. We follow recursively this procedure, until we eventually reach S_k, which finishes the classification procedure. Note that we adjust C_k to include only L_i labels.

[1] Available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

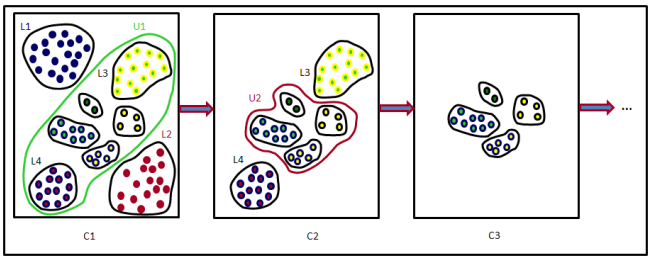

Fig. 3. Illustration of second ensemble SVM scheme for $m = 2$. The first cluster, C_1, consists of classes $\{L_1, L_2, U_1\}$. The second cluster, C_2, consists of $\{L_3, L_4, U_2\}$, and so on.

3 Experimental Setup

3.1 Image Data

The database released for the ImageCLEF-2009 Medical Annotation task includes 12677 fully classified (2D) radiographs for training and a separate test set consisting of 2000 radiographs. The aim is to automatically classify the test set using four different label sets including 57 to 193 distinct classes. A more detailed explanation of the database and the tasks can be found in [10].

3.2 Evaluation

We evaluate our SVM-based learning using two schemes depending on the availability of test data labels: 1)5-fold cross validation if test data labels are missing, and 2)ImageCLEF error counting scheme, otherwise. In the former scheme, the training database is partitioned into five subsets. Each subset is used once for testing while the rest are used for training, and the final result is assigned as the average of the five validations. Note that for each validation all classes were equally divided among the folds. We measure the overall classification performance using accuracy, which is the number of correct predictions divided by the total number of images. To the contrary, the error counting scheme is introduced by the challenge organizers to compare all runs submitted. Further details on this scheme can be found in [10].

3.3 Runs Submitted

As Computer Vision and Pattern Analysis (VPA) Laboratory of Sabanci University, we submitted three different runs to the ImageCLEF 2009 medical image annotation task. One obtained by the direct scheme (VPA-SABANCI-1), and two with the ensemble schemes (VPA-SABANCI-2 and -3). For each run, the optimum parameter setting was realized by trial-and-error.

4 Results

In this section, we present the results obtained by the proposed annotation schemes. In Table 1 we observe the results realized on the training database with 5-fold cross-validation. Ensemble scheme based on IRMA sub-codes clearly outperforms others, especially in terms of the 2007, 2008 and overall accuracies.

		Accuracy (%)				
Run	Type	2005	2006	2007	2008	Average
VPA-SABANCI-1	D	88.0	83.2	83.2	83.1	84.4
VPA-SABANCI-2	E-1	88.0	83.2	91.7	93.0	89.0
VPA-SABANCI-3	E-2	83.3	77.4	77.6	77.6	79.0

Table 1. Performance of VPA-SABANCI runs on training data.

Table 2 provides a detailed performance comparison of the direct scheme and the IRMA sub-codes based ensemble one over 2007 and 2008 labels. Simplifying the classification task by training a separate SVM over each sub-code, considerably improves the final accuracy relative to the usage of a single SVM. Furthermore, 2008 accuracies of individual SVMs excel those of 2007 despite higher number of classes (thus a more difficult classification problem). The underlying reason for this observation may be attributed to the more realistic labels of 2008.

	Ensemble by IRMA sub-codes					Direct
	SVM_1	SVM_2	SVM_3	SVM_4	Final	-
2007 accuracy (%)	96.7(5)	85.6(27)	88.0(66)	96.4(6)	**91.7**	**83.2**
2008 accuracy (%)	99.2(6)	86.3(34)	88.0(97)	98.5(11)	**93.0**	**83.1**

Table 2. Efficacy of ensemble classification based on IRMA sub-codes. Values in parenthesis refer to the number of distinct classes for that sub-task.

The results achieved on the test dataset in terms of prediction errors are presented in Table 3, together with the results of the best run realized in the challenge for comparison. As observed, IRMA sub-codes based ensemble scheme (E-1) outperforms its rivals again. With this performance, VPA-SABANCI-2 run is ranked 7^{th} among 18 runs submitted to the competition. Compared to our solution, the best run of the challenge exploits multiresolution analysis.

Figure 4 displays exemplary confusions realized by the best performing VPA-SABANCI-2 run for a class with few samples that lead to low recognition performance (19,5%), which may be partly due to low number of examples, and partly because of high visual similarity between the confused classes and the reference class (Most confusions are between images of the same body part, i.e.

		Error				
Run	Type	2005	2006	2007	2008	Sum
VPA-SABANCI-1	D	578	462	201.31	272.61	1513.92
VPA-SABANCI-2	E-1	578	462	155.05	261.16	1456.21
VPA-SABANCI-3	E-2	587	498	169.33	300.44	1554.77
TAUbiomed (best run)		356	263	64.30	169.50	852.80

Table 3. Performance of VPA-SABANCI runs, in comparison with the best run of the challenge, on test data. D refers to direct scheme, while E-1 and E-2 refer to ensemble schemes based on IRMA code and data distribution, respectively.

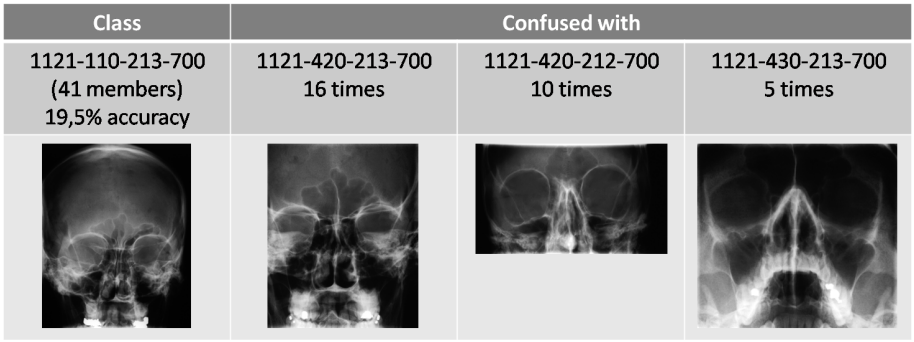

Fig. 4. Exemplary confusions realized by the proposed approach for a class with relatively low accuracy. Reference class with the corresponding label, number-of-examples, accuracy, and a representative X-ray image are shown on the left, while three most-observed confusions in descending order are displayed to the right.

the head. Note that, at manual categorization these images were assigned to different labels because of variation in image acquisition, such as view angle).

Table 4 demonstrates the computational requirements of the proposed schemes for testing. As observed, ensemble schemes require over 4-fold resources than the direct scheme on a single processor architecture. Nevertheless, this additional requirement can be canceled out by parallel processing.

5 Conclusion

In this paper we have introduced a classification work with the aim of automatically annotating X-Ray images. We have explored the annotation performances of two ensemble classification schemes based on individual SVMs trained on IRMA sub-codes and frequency of classes, and compared the results with the popular single-classifier scheme. Our experiments on the ImageCLEF-2009 Medical Annotation database revealed that breaking the annotation problem down to sub-problems by training individual SVMs over each IRMA sub-code outper-

Run	Type	CPU Time	Memory Usage
VPA-SABANCI-1	D	T	M
VPA-SABANCI-2	E-1	$4T$	M
VPA-SABANCI-3	E-2	kT	M

Table 4. Computational expense of VPA-SABANCI runs for testing on a PC with 2.40GHz processor and 6GB RAM. $T = 1.83$min, $M = 140$MB, and $k = \frac{\#classes}{m}$ with m being the split parameter defined in Section 2.2. Typically, $k > 4$ in our case.

forms its rivals in terms of annotation accuracy with the compromise of increased computational expense.

References

1. Long, L.R., Pillemer, S.R., Lawrence, R.C., Goh, G.H., Neve, L., Thoma, G.R.: WebMIRS: web-based medical information retrieval system. In Sethi, I.K., Jain, R.C., eds.: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Volume 3312. (December 1997) 392–403
2. Shyu, C.R., Brodley, C.E., Kak, A.C., Kosaka, A., Aisen, A.M., Broderick, L.S.: Assert: a physician-in-the-loop content-based retrieval system for hrct image databases. Comput. Vis. Image Underst. **75**(1-2) (1999) 111–132
3. Rahman, M.M., Desai, B.C., Bhattacharya, P.: Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. Computerized Medical Imaging and Graphics **32**(2) (2008) 95 – 108
4. Mueen, A., Sapian Baba, M., Zainuddin, R.: Multilevel feature extraction and x-ray image classification. J. Applied Sciences **7**(8) (2007) 1224–1229
5. Jacquet, V., Jeanne, V., Unay, D.: Automatic detection of body parts in x-ray images. In: Mathematical Methods in Biomedical Image Analysis, 2009. MMBIA 2009. IEEE Computer Society Workshop on. (2009)
6. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(7) (2002) 971–987
7. Unay, D., Soldea, O., Ekin, A., Cetin, M., Ercil, A.: Automatic Annotation of X-ray Images: A Study on Attribute Selection. In: Medical Content-based Retrieval for Clinical Decision Support (MCBR-CDS) Workshop in conjunction with MICCAI'09. (2009)
8. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2) (1998) 121–167
9. Müller, H., Deselaers, T., Deserno, T., Clough, P., Kim, E., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval. (2007) 595–608
10. Tommasi, T., Caputo, B., Welter, P., Güld, M.O., Deserno, T.M.: Overview of the clef 2009 medical image annotation task. In: Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers. Lecture Notes in Computer Science, Springer (2010)