ImageCLEF 2009 Medical Image Annotation Task: PCTs for Hierarchical Multi-Label Classification

Ivica Dimitrovski[1,2], Dragi Kocev[1], Suzana Loskovska[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
[2] Department of Computer Science, Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
{ivicad,suze}@feit.ukim.edu.mk,
{dragi.kocev,saso.dzeroski}@ijs.si

Abstract. In this paper, we describe an approach to the automatic medical image annotation task of the 2009 CLEF cross-language image retrieval campaign (ImageCLEF). This work focuses on the process of feature extraction from radiological images and their hierarchical multi-label classification. To extract features from the images we use two different techniques: edge histogram descriptor (EHD) and Scale Invariant Feature Transform (SIFT) histogram. To annotate the images, we use predictive clustering trees (PCTs) which are able to handle target concepts that are organized in a hierarchy, i.e., perform hierarchical multi-label classification. Furthermore, we construct ensembles (Bagging and Random Forests) that use PCTs as base classifiers: this improves the predictive/classification performance.

1 Introduction

The amount of medical images produced is constantly growing. Manual description and annotation of each image is time consuming, expensive and impractical. This calls for development of image annotation algorithms that can perform the task reliably. Automatic annotation classifies an image into one of a set of classes. If the classes are organized in a hierarchy and several of them can be assigned to an image, we are talking about hierarchical multi-label classification (HMLC).

This paper describes our approach to the medical image annotation task of ImageCLEF 2009 (for details see [1]). The objective of this task is to provide the IRMA (Image Retrieval in Medical Applications) code [2] for each image of a given set of previously unseen medical (radiological) images. The IRMA coding system consists of four axes: technical axis (T, image modality), directional axis (D, body orientation), anatomical axis (A, body region examined) and biological axis (B, biological system examined). The database of medical images contains 12677 fully annotated radiographs (training dataset for the classifier) and 1733 testing images without labels. The annotation should be performed by using the four different annotation label sets (the competitions from 2005-2008) in turn.

The code is strictly hierarchical because each sub-code element is connected to only one code element. This characteristic of the IRMA code allow us to exploit the

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part II, LNCS 6242, pp. 231–238, 2010.

code hierarchy and construct an automatic annotation system based on predictive clustering trees for hierarchical multi-label classification [3]. This approach is directly applicable for the datasets of ImageCLEF2007 and ImageCLEF2008 where the images were labeled according to the IRMA code scheme. To apply the same algorithm for the ImageCLEF2005 and ImageCLEF2006 datasets, we mapped the class numbers with the corresponding IRMA codes. Some images from the ImageCLEF2005 dataset can belong to more than one IRMA code. In the classification process, we use the most general IRMA code (that contains 0) to describe these images.

Automatic image classification/annotation relies on numerical features that are computed from the image pixel values. In our approach, we use an edge histogram descriptor (to extract the global features of the images) and SIFT histogram (to extract the local features from the images). We combine the feature vectors (histograms) with simple concatenation in a single vector with 2080 features.

The purpose of the concatenation of the global and the local features is to tackle the problem of intra-class variability vs. inter-class similarity and the different distribution of images between the training and the testing dataset (the testing dataset contains many images of some classes that are under-represented in the training set). Tomassi et al. [4] show that high and mid level combination of the different feature extraction techniques yield better results when SVMs are used as classifiers. In our work, we use ensembles of predictive clustering trees [3,5]. The ensembles of trees, such as random forests, can effectively exploit the information provided by the large number of features. Thus, we expect that concatenation of the feature extraction techniques yields better performance than the other combination methods.

The remainder of the paper is organized as follows: Section 2 describes the techniques for feature extraction from images. Section 3 introduces predictive clustering trees and their use for HMLC. In Section 4, we explain the experimental setup. Section 5 reports the obtained results. Conclusions and a summary are given in Section 6, where we also discuss some directions for further work.

2 Feature Extraction from Images

This section describes the techniques for feature extraction from images that we use to describe the X-ray images from ImageCLEF 2009. We shortly describe the edge histogram descriptor and the scale invariant feature transform. To learn a classifier and to annotate the images from the testing set, we use the feature vector obtained with simple concatenation of the features obtained from these two techniques.

Edge Histogram Descriptor: Edge detection is a fundamental problem of computer vision and has been widely investigated [6]. The goal of edge detection is to mark the points in a digital image at which the luminous intensity changes sharply. An edge representation of an image drastically reduces the amount of data to be processed, yet it retains important information about the shapes of objects in the scene. Edges in images constitute important features to represent their content.

One way of representing important edge features is to use a histogram. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. To represent it, MPEG-7 contains edge histogram

descriptors (EHD). These basically represent the distribution of five types of edges in each local area called a sub-image. The sub-images are defined by dividing the image space into 4×4 non-overlapping blocks. Thus, the image partition always yields 16 equal-sized sub-images, regardless of the size of the original image.

To characterize the sub-images, we then generate a histogram of edge distribution for each sub-image. Edges in the sub-images are categorized into five types: vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. Thus, the histogram for each sub-image represents the relative frequency of occurrence of the five types of edges in the corresponding sub-image.

As a result, each local histogram contains five bins. Each bin corresponds to one of the five edge types. Since there are 16 sub-images in the image, a total of 5×16=80 histogram bins are required. Note that each of the 80-histogram bins has its own semantics in terms of location and edge type. Edge detection is performed using the Canny edge detection algorithm [7].

SIFT histogram: Many different techniques for detecting and describing local image regions have been developed [8]. The Scale Invariant Feature Transform (SIFT) was proposed as a method of extracting and describing key-points which are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint [8].

For content based image retrieval, good response times are required and this is hard to achieve when using the huge amount of data contained in descriptors by local features. The descriptors using local features can be extremely big because an image may contain many key-points, each described by a 128 dimensional vector. To reduce the descriptor size, we use histograms of local features [9]. With this approach, the amount of data is reduced by estimating the distribution of local feature values for every image.

The creation of these histograms is a three step procedure. First, the key-points are extracted from all database images, where a key-point is described with a 128 dimensional vector of numerical values. For the key-point extraction and descriptor calculation, we use the default parameters proposed by Lowe [8]. The key-points are clustered in 2000 clusters using k-means. Afterwards, for each key-point we discard all information except the identifier of the most similar cluster center. A histogram of the occurring patch-cluster identifiers is created for each image. To be independent of the total number of key-points in an image, the histogram bins are normalized to sum to 1. This results in a 2000 dimensional histogram.

3 Ensembles of PCTs

In this section, we discuss the approach we use to classify the data at hand. We shortly describe the predictive clustering trees (PCT) framework, its use for HMLC and the learning of ensembles.

PCTs for Hierarchical-Multi Label Classification: In the PCT framework [5], a tree is viewed as a hierarchy of clusters: the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. PCTs can be constructed with a standard "top-down induction of

decision trees" (TDIDT) algorithm. The heuristic for selecting the tests is the reduction in variance caused by partitioning the instances. Maximizing the variance reduction maximizes cluster homogeneity and improves predictive performance.

A leaf of a PCT is labeled with/predicts the prototype of the set of examples belonging to it. With instantiation of the variance and prototype functions, the PCTs can handle different types of data, e.g., multiple targets [10] or time series [11]. A detailed description of the PCT framework can be found in [5].

To apply PCTs to the task of HMLC the example labels are represented as vectors with Boolean components. The *i*-th component of the vector is 1 if the example belongs to class c_i and 0 otherwise (See Fig. 1). The variance of a set of examples (S) is defined as the average squared distance between each example's label v_i and the mean label $\bar{v}$ of the set, i.e.,

$$Var(S) = \frac{\sum_i d(v_i, \bar{v})^2}{|S|} \tag{1}$$

The higher levels of the hierarchy are more important: an error in the upper levels costs more than an error on the lower levels. Considering that, a weighted Euclidean distance is used as a distance measure.

$$d(v_1, v_2) = \sqrt{\sum_i w(c_i)(v_{1,i} - v_{2,i})^2} \tag{2}$$

where $v_{k,i}$ is the i'th component of the class vector v_k of an instance x_k, and the class weights $w(c)$ decrease with the depth of the class in the hierarchy. In the case of HMLC, the notion of majority class does not apply in a straightforward manner. Each leaf in the tree stores the mean $\bar{v}$ of the vectors of the examples that are sorted in that leaf. Each component of $\bar{v}$ is the proportion of examples $\bar{v}_i$ in the leaf that belong to class c_i. An example arriving in the leaf can be predicted to belong to class c_i if $\bar{v}_i$ is above some threshold t_i. The threshold can be chosen by a domain expert. A detailed description of PCTs for HMLC can be found in [3].

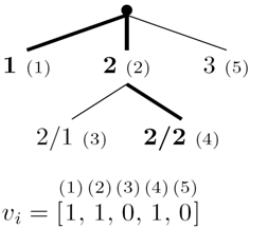

Fig. 1. A toy hierarchy. Class label names reflect the position in the hierarchy, e.g., '2/1' is a subclass of '2'. The set of classes {1, 2, 2/2} is indicated in bold in the hierarchy and is represented as a vector.

Ensemble Methods: An ensemble classifier is a set of classifiers. Each new example is classified by combining the predictions of every classifier from the ensemble.

These predictions can be combined by taking the average (for regression tasks) or the majority vote (for classification tasks) [12,13], or by taking more complex combinations. We have adopted the PCTs for HMLC as base classifiers. Average is applied to combine the predictions of the different trees because the leaf's prototype is the proportion of examples of different classes that belong to it. Just like for the base classifiers a threshold should be specified to make a prediction. We consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests.

Bagging [12] constructs the different classifiers by making bootstrap replicates of the training set and using each of these replicates to construct one classifier. Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set, until a number of instances is obtained equal to the size of the training set. Bagging is applicable to any type of learning algorithm.

A random forest [13] is an ensemble of trees, where diversity among the predictors is obtained both by bootstrap sampling, and by changing the feature set during learning. More precisely, at each node in the decision tree, a random subset of the input attributes is taken, and the best feature is selected from this subset (instead of the set of all attributes). The number of attributes that are retained is given by a function f of the total number of input attributes x (e.g., $f(x)=1$, $f(x)=\sqrt{x}$, $f(x)=\lfloor \log_2 x \rfloor+1$, ...). By setting $f(x)=x$, we obtain the bagging procedure. PCTs for HMLC are used as base classifiers.

4 Experimental Design

We decided to split the training images into training and development images. To tune the system for different distribution of images across classes in the training set and the test set, we generated several splits where the distributions of the images differed (in varying ways) between the training and development data.

We constructed a classifier for each axis from the IRMA code separately (see Section 1). From each of the datasets, we learn a PCT for HMLC and Ensembles of PCTs (Bagging and Random Forests). The ensembles consisted of 100 un-pruned trees. The feature subset size for Random Forests was set to 11 (using the formula $f(2080)=\lfloor \log_2(2080) \rfloor$).

To compare the performance of a single tree and an ensemble we use Precision-Recall (PR) curves. These curves are obtained with varying the value for the classification threshold: a given threshold corresponds to a single point from the PR-curve. For more information, see [3].

According to these experiments and previous research the ensembles of PCTs have higher performance as compared to a single PCT when used for hierarchical annotation of medical images [14]. Furthermore, the Bagging and Random Forest methods give similar results. Because the Random Forest method is much faster than the Bagging method, we submitted only the results for the Random Forest method.

To select an optimal value of the threshold (t), we performed validation on the different development sets. The threshold values that give the best results were used for the prediction of the unlabelled radiographs according to the four different classification schemes (see Section 1).

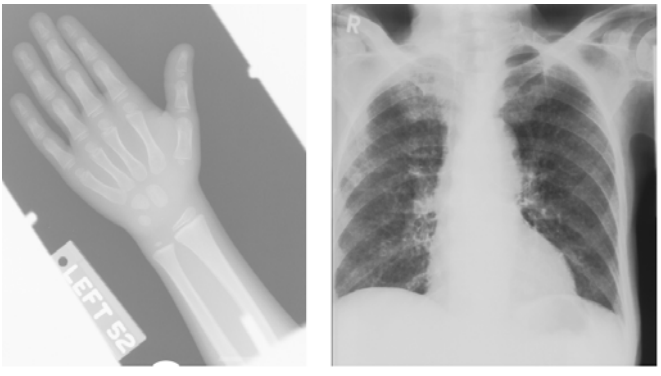

Fig. 2. Example images with same value for axis D, but different values for the axis combining D with the first code from A

To reduce the intra-class variability for axis D and improve the prediction performance, we decided to modify the hierarchy for this axis and include the first code of axis A from the corresponding IRMA code. Fig. 2 presents example images that have the same code for axis D, but are visually very different. After inclusion of the first code from the axis A, these images belong to different classes.

5 Results

For the ImageCLEF 2009 medical annotation task, we submitted one run. In this task, our result was third among the participating groups, with a total error score of 1352.56. The results for the particular datasets are presented in Table 1. From the results, we can note the high error for the annotations from ImageCLEF2005 and ImageCLEF2006. Recall that we pre-processed the images and the classes from 2005 and 2006 were mapped to an IRMA code. One class from the annotation from ImageCLEF2005 corresponds to multiple labels from the hierarchical annotation of the IRMA code and we used the most general class. This restricted the classifier to make more specific predictions. The performance for the ImageCLEF2008 is worse than the performance for ImageCLEF2007 because ImageCLEF2008 has a bigger hierarchy and more test images.

Similar conclusions can be made by analyzing the PR curves shown in Fig. 3. For each of the axes (T, D, A and B) we present three PR curves that correspond to the different annotation schemes. The PR curves for 2006 and 2007 coding schemes are equal because we simply mapped the class numbers to the corresponding IRMA codes. From the presented values for the $AU\overline{PRC}$ (Area under the Average Precision-Recall Curve) it can be seen that we obtain best results for the ImageCLEF2007 dataset. The $AU\overline{PRC}$ values for the ImageCLEF2005 dataset are very low considering the total number of classes, but this is mainly because we didn't apply a one-to-one mapping as for the ImageCLEF2006 dataset.

Table 1. Error score for the medical image annotation task and $AU\overline{PRC}$ per axis, using random forests of PCTs for HMLC

Annotation label sets	Error score	Number of wildcards (*)	$AU\overline{PRC}$ / RF			
			Axis T	Axis D	Axis A	Axis B
2005	549	0	0.9990	0.7712	0.7059	0.9843
2006	433	0	0.9998	0.8177	0.7419	0.9948
2007	128.1	2550	0.9998	0.8177	0.7419	0.9948
2008	242.26	2613	0.9995	0.7488	0.6621	0.9760

The excellent performance for the prediction task for axes T and B is due to the simplicity of the problem, the hierarchies along these axes contain only a few nodes (8 and 19 nodes for ImageCLEF2008, respectively). This means that in each node in the hierarchy there is a large portion of the examples, thus learning a good classifier is not a difficult task. The classifiers for the other two axes have satisfactory predictive performance, but here the predictive task is somewhat more difficult (especially for axis A). The size of the hierarchy along the A and D axis, for ImageCLEF2008 are 202 and 88 nodes, respectively.

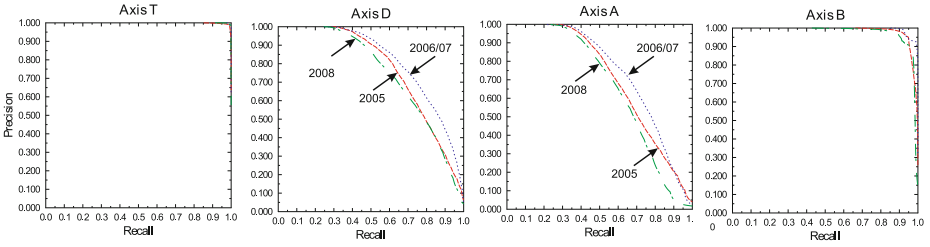

Fig. 3. Precision-Recall curves for the random forest predictions of the codes for T, D, A and B axis, respectively, for the four different competition tasks. The PR curves for the axes T and B are close to each other for each year. For the axes D and A, the upper PR curves are for the years 2006/07, the lower ones are for 2008 and the PR curves in the middle are for 2005.

6 Conclusions

This paper presents a hierarchical multi-label classification approach to medical image annotation. For efficient image representation, we use edge histogram descriptor and SIFT histograms. The predictive modeling problem that we consider is to learn PCTs and ensembles of PCTs that predict a hierarchical annotation of an X-ray image. Using these approaches, we obtained good predictive performance and ranked third on the ImageCLEF 2009 competition.

There are several ways to further improve the predictive performance of the proposed approach. First, one could try to tackle the shift in distribution of images between the training and the testing set. One solution is to develop extensions of the PCT approach that can handle such differences. Another approach is to generate virtual samples of the images that are underrepresented in the training set by rotation,

translation and manipulation of contrast and brightness. Second, better performance may be obtained by post-processing the output from the ensembles and by reducing the dependence from the thresholding: instead of the hard threshold, use the raw probabilities. Third, we could use additional feature extraction techniques and combine them using different combination schemes (other than concatenation).

In summary, we presented a general approach to hierarchical image annotation. The approach can be easily extended with new feature extraction methods, and can thus be applied to other domains. It can be also easily applied to arbitrary domains, because it can handle hierarchies with arbitrary sizes (bigger hierarchies, hierarchies that are organized as trees or directed acyclic graphs).

References

1. Tommasi, T., Caputo, B., Welter, P., Guld, M.O., Deserno, T.M.: Overview of the CLEF 2009 medical image annotation track. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part II. LNCS, vol. 6242, Springer, Heidelberg (2010)
2. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: Proc. of SPIE - Medical Imaging 2003, vol. 5033, pp. 440–451 (2003)
3. Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning 73(2), 185–214 (2008)
4. Tommasi, T., Orabona, F., Caputo, B.: Discriminative cue integration for medical image annotation. Pattern Recognition Letters 29(15), 1996–2002 (2008)
5. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Proc. of the 15th ICML, pp. 55–63 (1998)
6. Ziou, D., Tabbone, S.: Edge Detection Techniques an Overview. International Journal of Pattern Recognition and Image Analysis 8(4), 537–559 (1998)
7. Canny, J.F.: A computational approach to edge detection. IEEE Trans. Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
8. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: CVPR 2005, San Diego, CA, vol. 2, pp. 157–162 (2005)
10. Kocev, D., Vens, C., Struyf, J., Dzeroski, S.: Ensembles of Multi-Objective Decision Trees. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 624–631. Springer, Heidelberg (2007)
11. Dzeroski, S., Gjorgjioski, V., Slavkov, I., Struyf, J.: Analysis of Time Series Data with Predictive Clustering Trees. In: Džeroski, S., Struyf, J. (eds.) KDID 2006. LNCS, vol. 4747, pp. 63–80. Springer, Heidelberg (2007)
12. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
13. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
14. Dimitrovski, I., Kocev, D., Loskovska, S., Dzeroski, S.: Hierarchical annotation of medical images. In: Proc. of the 11th International Multiconference – IS 2008, Ljubljana, Slovenia, pp. 170–174 (2008)