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

Ivica Dimitrovski[1], Dragi Kocev[2], Suzana Loskovska[1], Saso Dzeroski[2]

[1] Department of Computer Science, Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia

[2] Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia

Abstract

In this paper, we describe an approach for the automatic medical image annotation task of the 2009 CLEF cross-language image retrieval campaign (ImageCLEF). This work is focused on the process of feature extraction from radiological images and hierarchical multi-label classification. To extract features from the images we used an edge histogram descriptor as global feature and SIFT histogram as local feature. These feature vectors were combined through simple concatenation in one feature vector with 2080 variables. With the combination of global and local features we want to tackle the problem of intra-class variability vs. inter-class similarity and the problem of data unbalance between train and test datasets. For classification we selected an extension of the predictive clustering trees (PCTs) able to handle data types organized in hierarchy. Furthermore, we constructed ensembles (Bagging and Random Forests) that use PCTs as base classifiers to improve the performance.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

Keywords

Automatic Image Annotation, Scale Invariant Feature Transform, Edge Histogram Descriptor, Hierarchical Multi-Label Classification, Predictive Clustering Trees, Random Forests

1 Introduction

The amount of medical images produced nowadays is constantly growing. Manual description and annotation of each image is time consuming, expensive and impractical. This calls for development of automatic image annotation algorithms that can perform the task reliably. With the automatic annotation an image is classified into set of classes. If these classes are organized in a hierarchy then it is a case of hierarchical multi-label classification.

This paper describes our approach for the medical image annotation task of ImageCLEF 2009. The objective of this task is to provide the IRMA (Image Retrieval in Medical Applications) code [1] for each image of a given set of previously unseen medical (radiological) images. The results of the classification step can be used for multilingual image annotations as well as for DICOM standard header corrections.

A database of 12677 fully classified radiographs, taken randomly from medical routine, is provided to be used in any way to train a classifier. Images are labeled according to four classification label sets considering:

- 57 classes as in ImageCLEF 2005
- 116 classes as in ImageCLEF 2006
- 116 IRMA codes as in ImageCLEF 2007
- 193 IRMA codes as in ImageCLEF 2008

In the test phase we have to classify and assigned the correct labels to 1733 radiographs according to the four different schemes.

The IRMA coding system consists of four axes with three to four positions, each in {0,…, 9, a,…, z}, where "0" denotes "unspecified" to determine the end of a path along an axis:

- T (Technical): image modality
- D (Directional): body orientation
- A (Anatomical): body region examined
- B (Biological): biological system examined

The code is strictly hierarchical because each sub-code element is connected to only one code element. The element to the right is a sub element of the element to the left. These characteristics of the IRMA code lead as to exploit the code hierarchy and construct an automatic annotation system based on predictive clustering trees framework for hierarchical multi-label classification [2]. This approach was directly applicable for the datasets of ImageCLEF 2007 and ImageCLEF 2008 because the images were labeled according to the IRMA code scheme. To apply the same algorithm for the ImageCLEF 2005 and ImageCLEF 2006 datasets we mapped the class numbers with the corresponding IRMA codes. The images from ImageCLEF 2005 dataset belong to more than one IRMA code. In the classification process we used the most general IRMA code to describe these images.

Automatic image classification relies on numerical features that are computed from the pixel values. In our approach we used an edge histogram descriptor as global feature and SIFT histogram as local feature. These feature vectors were combined through simple concatenation in one feature vector with 2080 variables. With the combination of global and local features we want to tackle the problem of intra-class variability vs. inter-class similarity and the problem of data unbalance between train and test datasets.

The rest of the paper is organized as follows: section 2 describes the feature extracted from images. Section 3 gives details on the predictive clustering trees framework and its use for hierarchical multi-label classification. In section 4 we explain the experimental setup. Section 5 reports the obtained results. Conclusions and summary are given in Section 6.

2 Feature Extraction from Images

This section describes the features we used to describe the X-ray images from ImageCLEF 2009. We shortly describe the edge histogram descriptor and scale invariant feature transform. In the classification phase we used the feature vector obtained with simple concatenation of these features.

2.1 Edge Histogram Descriptor

Edge detection is a fundamental problem of computer vision and has been widely investigated [3]. The goal of edge detection is to mark the points in a digital image at which the luminous intensity changes sharply. 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 an important feature to represent their content. One way of representing such an important edge feature 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 this unique feature, in MPEG-7, there is a edge histogram descriptor (EHD) in the image. The EHD basically represents the distribution of five types of edges in each local area called a sub-image. As shown in Figure 1, the sub-image is defined by dividing the image space into 4×4 nonoverlapping blocks. Thus, the image partition always yields 16 equal-sized sub-images regardless of the size of the original image.

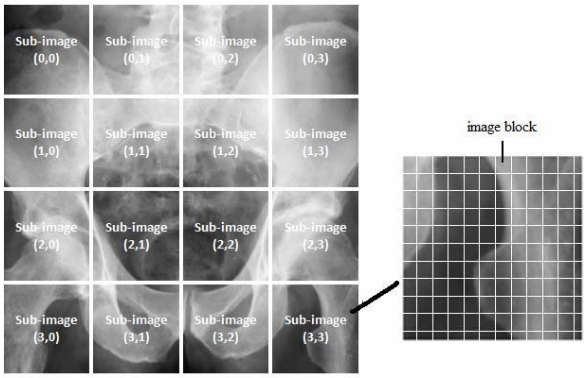

Figure 1: Definition of sub-image and image-block.

To characterize the sub-image, 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 five types of edges in the corresponding sub-image.

As a result, each local histogram contains five bins. Each bin corresponds to one of five edge types. Since there are 16 sub-images in the image, a total of 5×16=80 histogram bins is required. Note that each of the 80-histogram bins has its own semantics in terms of location and edge type. For example, the bin for the horizontal type edge in the sub-image located at (0,0) in Figure 1 carries the information of the relative population of the horizontal edges in the top-left local region of the image. The edge detection was performed using Canny edge detection algorithm [4].

Because of the low contrast of the X-ray images we applied a contrast enhancement technique for the images used in our experiments. The contrast enhancement was done through histogram equalization for the central part of the images, because the image corners have only black pixels.

2.2 SIFT histogram

Many different techniques for detecting and describing local image regions have been developed [5]. The Scale Invariant Feature Transform (SIFT) was proposed as a method of extracting and describing keypoints which are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint [5].

For content based image retrieval good response times are required and this is hard to achieve using the huge amount of data obtained by local features. The dimension of this feature is extremely high because the size of the keypoint descriptor is 128 dimensional vector. To reduce the dimensionality we use histograms of local features [6]. With this approach the amount of data is reduced by estimating the distribution of local features 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 vector of numerical values. The key-points are then clustered in 2000 clusters. 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. This results in a 2000 dimensional histogram per image.

3 Ensembles for PCTs

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

3.1 PCTs for Hierarchical-Multi Label Classification

In the PCT framework [7], 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. With instantiation of the variance and prototype function the PCTs can handle different types of data, e.g. multiple targets [8] or time series [9]. A detailed description of the PCT framework can be found in [7].

To apply PCTs to the task of HMLC first 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 Figure 2). Then 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.,

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

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, weighted Euclidean distance is used as a distance measure.

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

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. Second, 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 the PCTs for HMLC can be found in [10].

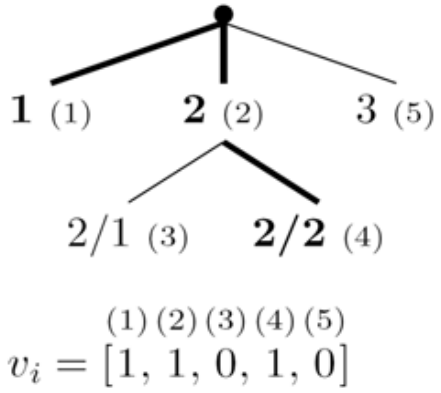

Figure 2: A toy hierarchy. Class label names reflect the position in the hierarchy, e.g., '2.1' is a subclass of '2'. The set of classes {1, 2, 2.2}, indicated in bold in the hierarchy, and represented as a vector.

3.2 Ensemble Methods

An ensemble is a set of classifiers constructed with a given algorithm. 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) [11], [12], or by taking more complex combinations. We consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests.

Bagging [11] 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 an equal number of instances is obtained.

A random forest [12] is an ensemble of trees, where diversity among the predictors is obtained by using bagging, and by changing the feature set during learning. More precisely, at each node in the decision trees, a random subset of the input attributes is taken, and the best feature is selected from this subset. The number of attributes that are retained is given by a function f of the total number of input attributes x (e.g., $f(x)=1$, $f(x)=\sqrt{x}$, $f(x)=\lfloor \log_2 x \rfloor +1, \ldots$). By setting $f(x)=x$, we obtain the bagging procedure. The PCTs for HMLC are used as base classifiers. Average is applied to combine the different predictions because the leaf's prototype is the proportion of examples that belong to it. This means that a threshold should be specified to make a prediction.

4 Experimental Design

We decided to split the training images into training and development images because the organizers didn't supply development set. To tune the system for different distribution of the images in the classes of the training set and the test set we generated several splits with different distribution of the images in the classes of the training and development data.

We decided to learn a classifier for each axis 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 threshold: a given threshold corresponds to a single point from the PR-curve. For more information, see [10]. According to these experiments and previous research the ensembles of PCTs showed increased performance as compared to a single PCT when applied for hierarchical annotation of medical images [13]. Furthermore, the Bagging and

Random Forest methods give similar results and 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).

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. Figure 3 presents example images that have same code for axis D but are visually different. After inclusion of the first code from the axis A these images belong to different classes.

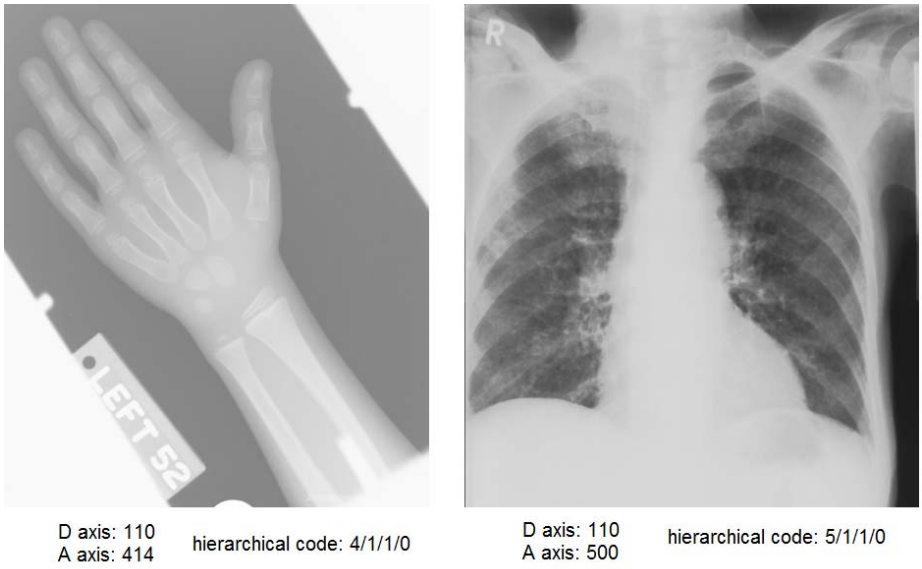

Figure 3: Example images with same value for axis D.

5 Results

For ImageCLEF 2009 medical annotation task we submitted one run. In this task our result was third among the participating groups, with total error score of 1352.56. The results for the particular datasets are shown in Table 1.

Table 1: Error score for the medical image annotation run

classification label sets	Error score
2005	549
2006	433
2007	128.1
2008	242.26

6 Conclusion

This paper presented a hierarchical multi-label classification approach for medical image annotation. For efficient image representation we used edge histogram descriptor and SIFT histogram. 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.

The proposed approach can be easily extended with new feature extraction methods, and can thus be applied to other domains. The proposed approach for hierarchical annotation can be 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] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, B. B. Wein, Berthold. The IRMA code for unique classification of medical images, Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation, Proceedings of the SPIE, Vol. 5033, pp. 440–451, 2003.

[2] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification, Machine Learning Journal 73(2), pp. 185-214, 2008.

[3] D. Ziou, S. Tabbone.Edge Detection Techniques An Overview, International Journal of Pattern Recognition and Image Analysis, 8(4), pp. 537-559, 1998.

[4] J.F. Canny. A computational approach to edge detection. IEEE Trans Pattern Analysis and Machine Intelligence, 8(6): 679-698, Nov 1986.

[5] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91–110, 2004.

[6] T. Deselaers, D. Keysers, H. Ney. Discriminative training for object recognition using image patches. In: CVPR 05. Volume 2., San Diego, CA (2005) 157–162.

[7] H. Blockeel, L. De Raedt and J. Ramon. Top-down induction of clustering trees. In Proc. of the 15th ICML, p.55-63, 1998.

[8] D.Kocev, C. Vens, J. Struyf, S. Dzeroski. Ensembles of Multi-Objective Decision Trees, In Proc. of the ECML 2007, LNAI vol. 4701, p. 624-631, 2007.

[9] S. Dzeroski, V. Gjorgjioski, I. Slavkov, J. Struyf. Analysis of Time Series Data with Predictive Clustering Trees, In KDID06, LNCS vol. 4747, p. 63-80, 2007.

[10] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, H. Blockeel. Decision trees for hierarchical multi-label classification, Machine Learning Journal, DOI-10.1007/s10994-008-5077-3, 2008.

[11] L. Breiman. Bagging predictors, Machine Learning Journal, vol. 24 Issue 2, p. 123-140, 1996

[12] L. Breiman. Random Forests, Machine Learning Journal, vol. 45, p.5-32, 2001.

[13] I. Dimitrovski, D. Kocev, S. Loskovska, S. Dzeroski, Hierarchical annotation of medical images, Proceedings of the 11th International Multiconference, Information Society − IS 2008, Data Mining and Data Warehouses, Oct. 2008 Ljubljana, Slovenia, pp. 170-174.