HIERARCHICAL ANNOTATION OF MEDICAL IMAGES

Ivica Dimitrovski[1]*, Dragi Kocev*[2]*, Suzana Loškovska*[1]*, Sašo Džeroski*[2]
[1]Department of Computer Science, Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia
e-mail: {ivicad, suze}@feit.ukim.edu.mk
[2]Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
e-mail: {Dragi.Kocev, Saso.Dzeroski}@ijs.si

ABSTRACT

In this paper, we describe an approach for the automatic medical annotation task of the 2008 CLEF cross-language image retrieval campaign (ImageCLEF). The data comprise 12076 fully annotated images according to the IRMA code. This work is focused on the process of feature extraction from images and hierarchical multi-label classification. To extract features from the images we used a technique called: local distribution of edges. With this techniques each image was described with 80 variables. The goal of the classification task was to classify an image according to the IRMA code. The IRMA code is organized hierarchically. Hence, as classifer we selected an extension of the predictive clustering trees (PCTs) that is able to handle this type of data. Further more, we constructed ensembles (Bagging and Random Forests) that use PCTs as base classifiers.

1 INTRODUCTION

The amount of medical images produced nowadays is constantly growing. The cost of manually annotating these images is very high. 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 the medical annotation task of ImageCLEF 2008 [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. 12,076 classified training images are provided to be used in any way to train a classifier. The results of the classification step can be used for multilingual image annotations as well as for DICOM standard header corrections. According to the IRMA code [2], a total of 197 classes are defined. 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

This allows a short and unambiguous notation (IRMA: TTTT-DDD-AAA-BBB), where T, D, A, and B denotes a coding or sub-coding digit of the respective axis. Figure 1 gives two examples of unambiguous image classification using the IRMA code. The image on the left is coded: **1123** (x-ray, projection radiography, analog, high energy) – **211** (sagittal, left lateral descubitus, inspiration) – **520** (chest, lung) – **3a0** (respiratory system, lung). The image of the right is coded: **1220** (x-ray, fluoroscopy, analog) – **127** (coronad, ap, supine) – **722** (abdomen, upper abdomen, middle) – **430** (gastrointestinal system, stomach).

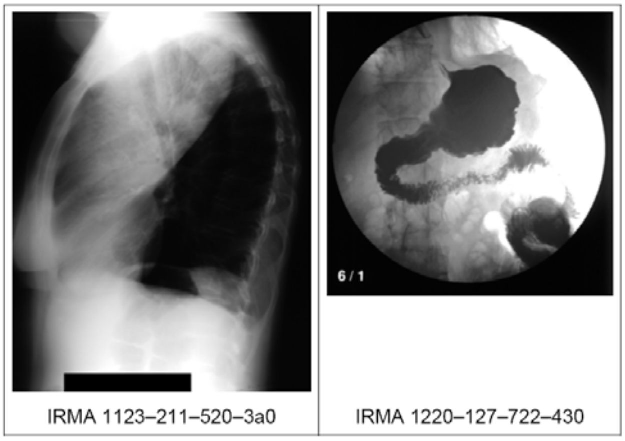

Figure 1: *IRMA-coded chest and abdomen radiograph.*

The code is strictly hierarchical – 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. For example:

2	cardiovascular system
21	cardiovascular system; heart
216	cardiovascular system; heart; aortic valve

The aortic valve is an element of the heart, which in turn is an element of the cardiovascular system.

The difference between ImageCLEF 2008 task and the tasks from previous years is the distribution of images. To encourage the exploitation of the class hierarchy, the images in the 2008 test set are mainly from classes which have only few examples of the same class in the training data and thus it is significantly harder to consider this task as a flat classification task as most of the successful techniques did in 2007 [3]. Instead, it is expected that exploiting the hierarchy will lead to large improvements.

Automatic image classification relies on numerical features that are computed from the pixel values [4]. In our approach we use edge histogram descriptor to represents the spatial distribution of five types of edges (four directional edges and one non-directional, see Fig. 3).

For the classification task, we applied predictive clustering trees (PCTs) that are instantiated for handling hierarchical multi-label classification (HMLC) and ensembles of PCTs.The results show the increase of predictive power when ensembles are used as a classifier.

2 FEATURE EXTRACTION FROM IMAGES: HISTOGRAM OF LOCAL EDGES DISTRIBUTION

Edge detection is a fundamental problem of computer vision and has been widely investigated [5]. 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 descriptor for edge distribution (EHD) in the image. The EHD basically represents the distribution of 5 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.

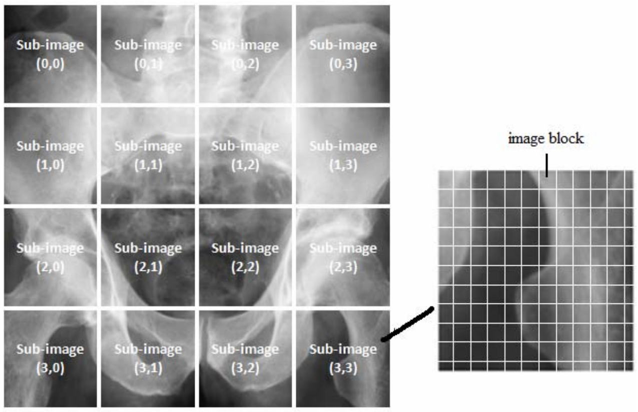

Figure 2: *Definition of sub-image and image-block.*

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 5 types: vertical, horizontal, 45-degree diagonal, 135-degree diagonal, and non-directional edges (see Figure 3). Thus, the histogram for each sub-image represents the relative frequency of occurrence of the 5 types of edges in the corresponding sub-image.

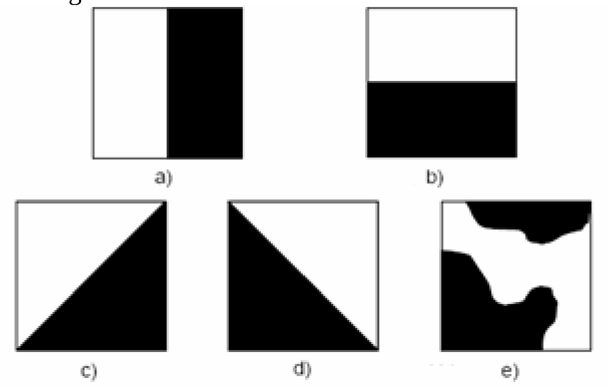

Figure 3: *Five types of edges: a) vertical edges, b) horizontal edge, c) 45-degree edge, d) 135-degree edge, e) non-directional edge*

As a result, each local histogram contains 5 bins. Each bin corresponds to one of 5 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 2 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 [6].

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.

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 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 that is used 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].

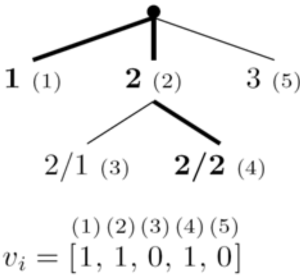

Figure 4: *A toy hierarchy. Class label names reflect the position in the hierarchy, e.g., '2.1' is a subclass of '2'. The set of classes {1, 2, 2.2}, indicated in bold in the hierarchy, and represented as a vector.*

In order to apply PCTs to the task of HMLC, the variance and prototype parameters were properly instantiated.

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 4). Then the variance of a set of examples (*S*) can be 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.

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

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 therefore be predicted to belong to class c_i if $\bar{v}_i$ is above some threshold t_i, which can be chosen by a domain expert. A detailed description of the PCTs for HMLC can be found in [10].

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.

In this paper, we consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests.

Bagging [11] is an ensemble method that 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 additionally 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., $f(x) = 1, f(x) = \sqrt{x}, f(x) = \lfloor \log_2 x \rfloor + 1, ...$). By setting $f(x) = x$, we obtain the bagging procedure.

In this work, the PCTs for HMLC are used as base classifiers. Average is applied to combine the different predictions. This is because the leaf's prototype is the proportion of examples that belong to it. This means that a threshold should be specified in order to make an prediction.

4 EXPERIMENTAL DESIGN

Here, we describe the setup we used to analyze the data.

For each of the axes (see the data description in Section 1) we have 4 training and 4 testing datasets. 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 7 (using the formula $f(80) = \lfloor \log_2 80 \rfloor + 1$).

To compare the performance of a single tree and an ensemble we use Precision-Recall (PR) curves (see Figure 5). 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].

To decide for an optimal value of the threshold (*t*), 10-fold cross validation on the training set is performed. From the PR curves one can select few thresholds and evaluate the predictions of the models for each of the threshold.

5 RESULTS AND DISCUSSION

The results from the experiments are shown in Figure 5. For each of the axes we present a PR curves for the three methods we use.

From the curves we can note the increase of the predictive performance when we use ensembles instead of single tree. The lift in performance that ensembles give to their base classifier was previously noted in the cases of classification and regression [11, 12] and multiple targets prediction [8].

The excellent performance for the prediction task for axes T and B (AUPRC of 0.9994 and 0.9862) is due to the simplicity of the problem. Namely, the hierarchies along these axes contain only few nodes (9 and 27, respectively). This means that in each node in the hierarchy there are nice portion of the examples, thus learning a good classifier is not a difficult task. The classifiers for the other two axes have high predictive performance (AUPRC of 9064 and 0.8264), but here the predictive task is somewhat more difficult (especially for axis A). The sizes of the hierarchies for axes A and D are 110 and 36 nodes, respectively.

A successfull image annotation system highly depends of the performance of its two main components: the feature extractor and the classifier. The feature extraction process should provide a vector of features that best reflects the different aspects for distinguishing one class from the others. When such features are given to a classifier that is able to capture the nature of the task, then the predictive performance of such a classifier will be very high.

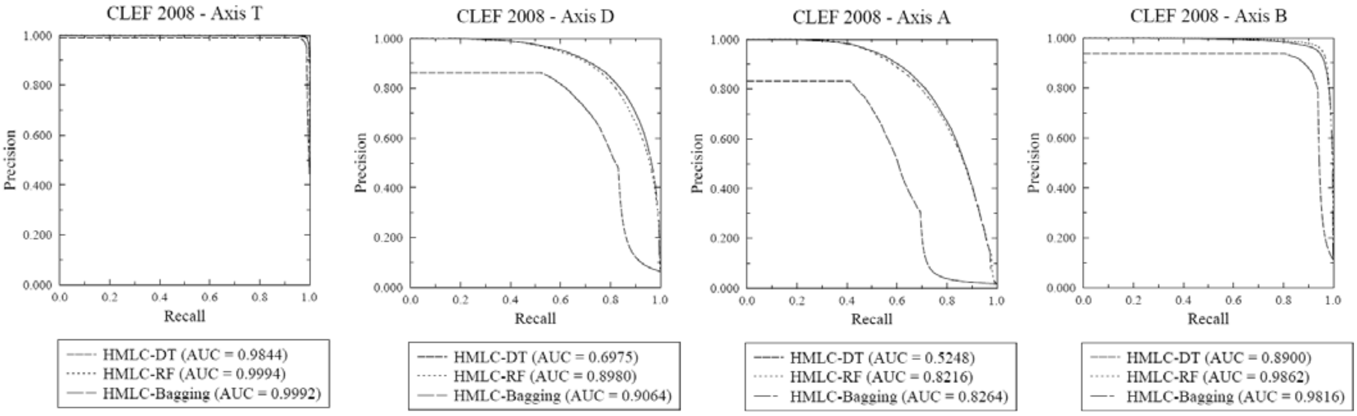

Figure 5: *Precision-Recall curves for the T,D,B and A axis, respectively*

6 CONCLUSIONS

This paper presented a hierarchical multi-label classification (HMLC) approach to medical image annotation. For efficient image representation we used local distribution of edges. The edge histogram is robust feature for representing gray-scale radiological images.

We applied PCTs for HMLC and ensembles of PCTs in order to accurately classify the image in the IRMA code hierarchy. The ensembles of PCTs showed increased performance as compared to a single PCT.

There are few possibilities for improvements of the results, that we plan to further investigate. First, we can further exploit the hierarchical nature of the IRMA code: instead of learning a classifier for each axis separately, one can learn a classifier for all the axes (or combinations of axes). Second, we plan to use other algorithms for feature extraction from images (e.g. Scale-invariant feature transform, SIFT) that were previously successfully used in image annotation [4].

Another line of further work is extensions of the machile learning algorithm. One such extension is enabling the algorithm to learn a model that ia aware of a covariate shift (the test set distribution is different from the train set distribution). Also, we plan to implement other distance measures for hierarchies (e.g. Jaccard similarity coefficient - like).

References

[1] http://www.imageclef.org/2008/medaat

[2] 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, Volume 5033, pp. 440-451, 2003

[3] H. Muller, T. Deselaers, E. Kim, J. Kalpathy–Cramer,T. M. Deserno, W. Hersh. Overview of the ImageCLEF 2007 Medical Retrieval and Annotation Tasks, Advances in Multilingual and Multimodal Information Retrieval 8th Workshop of the CLEF 2007, vol. 5152, Budapest, Hungary, Springer, 2007

[4] T. Deselaers, D. Keysers, H. Ney.Features for Image Retrieval: An Experimental Comparison, Information Retrieval, vol. 11, issue 2, The Netherlands, Springer, pp. 77-107, 2008

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

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

[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