Global and Local Multi-valued Dissimilarity-Based Classification: Application to Computer-Aided Detection of Tuberculosis

Yulia Arzhaeva[1,2], Laurens Hogeweg[1], Pim A. de Jong[1], Max A. Viergever[1], and Bram van Ginneken[1]

[1] Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[2] CSIRO Mathematical and Information Sciences, Australia
`yulia.arzhaeva@csiro.au`

Abstract. In many applications of computer-aided detection (CAD) it is not possible to precisely localize lesions or affected areas in images that are known to be abnormal. In this paper a novel approach to computer-aided detection is presented that can deal effectively with such weakly labeled data. Our approach is based on multi-valued dissimilarity measures that retain more information about underlying local image features than single-valued dissimilarities. We show how this approach can be extended by applying it locally as well as globally, and by merging the local and global classification results into an overall opinion about the image to be classified. The framework is applied to the detection of tuberculosis (TB) in chest radiographs. This is the first study to apply a CAD system to a large database of digital chest radiographs obtained from a TB screening program, including normal cases, suspect cases and cases with proven TB. The global dissimilarity approach achieved an area under the ROC curve of 0.81. The combination of local and global classifications increased this value to 0.83.

1 Introduction

Pulmonary tuberculosis (TB) is a major cause of death and illness worldwide, with 9.2 million new cases and 1.7 million deaths reported in 2006 [1]. Chest radiography is increasingly important in the fight against TB, especially because the rates of sputum-negative TB are rapidly increasing in populations with a high incidence of HIV/AIDS. On chest radiographs, TB often presents itself through subtle diffuse textural abnormalities. With the advent of digital radiography, computer-aided detection (CAD) systems can be developed that could facilitate mass population TB screening.

However, little research has been done in this area. In [2] texture analysis within the lung fields was used but this required experts to manually delineate abnormal areas, in order to train the system to discern normal regions from abnormal. Although such an approach may lead to a powerful CAD system, obtaining manual delineations of ill-defined diffuse lesions is laborious and likely to produce an unreliable ground truth. Our work focuses on classification of weakly labeled images,

i.e. when the exact locations of abnormalities in training data are unknown and, therefore, local feature-based classifiers cannot be trained. Our approach circumvents the problem of the absence of local ground truth by using the distances, or dissimilarities, between estimated distributions of local features in the global classification of images. We estimate these differences per feature and therefore build a multi-valued dissimilarity-based (MVDB) classification system.

The underlying assumption of the MVDB method is that local feature distributions are sufficiently different for normal and abnormal images. However, this assumption is not likely to hold for cases with subtle small abnormalities only. We hypothesize that subdividing the lung fields into smaller parts and applying the MVDB classification to these parts separately, and subsequently combining these local opinions, may improve the sensitivity of the method to such abnormalities and increase overall classification performance. It should be noted that obtaining ground truth labels for fixed large lung subdivisions is easier than obtaining manual delineations of lesions. In this work, we apply the method to a large database of digital radiographs from a TB screening program. The proposed modification of the MVDB classification is general and applicable to other image classification tasks that involve local analysis.

2 Methods

2.1 Multi-valued Dissimilarity-Based Classification

Dissimilarity-based classification uses dissimilarity representations of objects instead of traditional feature vectors, that is, objects are represented by their pairwise comparisons. This is a natural way to describe a class of similar objects. A pairwise comparison is done by computing a measure of dissimilarity, or distance, between two objects. In the standard dissimilarity-based classification [3], each training object is represented as a vector of distances to a set of prototype objects. Then, any traditional classifier can be trained on dissimilarity representations of training objects and applied to the dissimilarity representation of a new object. This may not be an efficient strategy for classifying objects characterized by a large set of descriptors, such as numerous local texture features, because it reduces the abundance of local information in two objects to just one dissimilarity value between them.

MVDB classification is built on similar principles but reduces the loss of information compared to standard dissimilarity-based classification. While the standard dissimilarity-based method accumulates the distance over all the object descriptors, the MVBD method is based on computing a distance for every descriptor individually.

Let x and y be two objects characterized by n one- or multi-dimensional descriptors f_i, and $d_i = d(f_i^x, f_i^y)$ be the value of dissimilarity between corresponding descriptors of x and y, where d is a dissimilarity measure. Then, a vector $D(x,y) = (d_1, \ldots, d_n)$ is called the dissimilarity representation of object x with respect to object y. To construct a classifier on such representations,

let us consider a training set T, and a set of prototype objects R of size r, $R = \{p_1, \ldots, p_r\}$, where $R \subseteq T$. For each $x \in T$, r different representations $D(x, p_k)$, $1 \le k \le r$, can be obtained, and consequently r classifiers can be trained on T using $D(x, p_k)$ as input. A test object, subsequently, can be classified r times using its prototype-bound representations. To obtain a final classification solution for a test object, the outputs of r classifiers must be combined. Combining classifiers benefits from complementary information provided by different dissimilarity representations. In this study we combine the posterior probabilities resulting from different classifiers with the sum rule:

$$P(c|x) = \frac{1}{r} \sum_{k=1}^{r} P_k(c|x) , \qquad (1)$$

where $P(c|x)$ is a posterior probability that the object x belongs to a class c, and $P_k(c|x)$ is a posterior probability yielded by the classifier k.

Figure 1 schematically depicts the steps of the MVDB classification. To apply this method to an image classification task that involve local texture analysis, we describe each image by the distributions of its local texture features. These features are extracted at numerous locations inside the image, and their individual distributions are estimated by histograms. From the set of training images, r prototype images are selected, either randomly, or by following a systematic approach. Since the sum rule, used to combine the results of individual classifiers, is known to be less sensitive than other combiners to the errors of individual classifiers [4], we believe, the random selection of prototypes is a reasonable starting approach. Dissimilarities, computed between corresponding feature histograms of the image and a prototype, constitute a dissimilarity representation of the image.
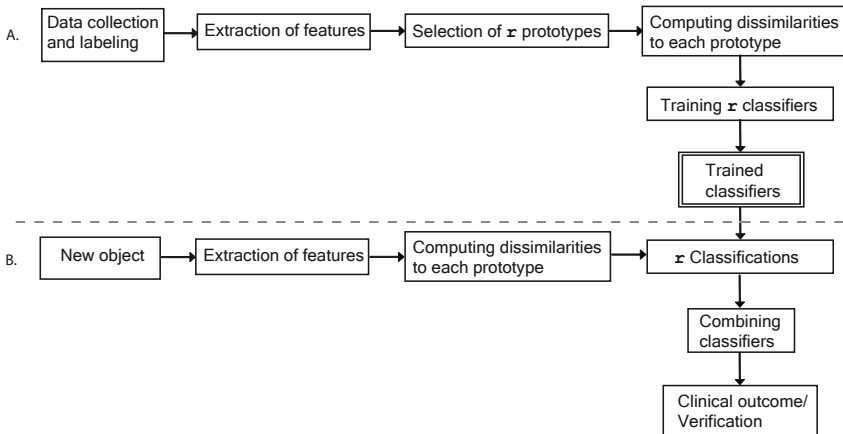
Fig. 1. Flow chart of the MVDB classification. (A) Training phase. (B) Testing phase.

2.2   Local Classification to Improve Global Results

Sometimes image descriptors, such as the local feature histograms, are too generalizing. This is true when an object whose presence we want to detect is too small with respect to the whole image. With the detection of TB, the texture feature histograms computed over the whole lung fields might not be sensitive enough to reflect the presence of subtle localized lesions in the lungs. We assume that the discriminating ability of descriptors will increase if they are computed over smaller image parts. When it is practical to obtain the ground truth for training images on a finer scale, e.g. class labels for a fixed image partition, we propose the following modification of the MVDB classification scheme.

1. Images are partitioned, and the ground truth is obtained for each part.
2. The MVDB scheme is applied to each image part separately, and, optionally, to the whole images too.
3. The classification results are combined to obtain an overall image solution.

Here, the combination rule might be different from the one in step 5 of the testing phase of the original scheme. In this paper we use the vote rule for the abnormal class ($c = 1$), and compute the posterior probability of the normal class($c = 0$) such as $P(c = 0|x) = 1 - P(c = 1|x)$, where x is the test object. The vote rule for computing $P(c = 1|x)$ is

$$P(c|x) = \max(P_0(c|x), \max_{l=1}^{L} P_l(c|x_l)), \ c = 1, \tag{2}$$

where x_l, $1 \leq l \leq L$, are L image subdivisions, P_l is the result of applying the MVDB classification to x_l, and P_0 is the result of applying the MVDB classification to the whole image. The choice for the vote rule for the detection of abnormal images is intuitive because if any part of the image is abnormal then the whole image is abnormal. The use of P_0 in Eq. 2 is optional and is not needed if the performances of all regional classifiers are considerably better than that of the global classifier. It should also be noted that, for a certain region, only a fraction of abnormal images will have abnormalities in that particular region. Therefore, a high classification performance on one of the regions is not enough to obtain an equally high performance after combining. In addition to improving the image classification performance, the application of the MVDB to the regions allows one to obtain a prediction on which regions are likely to contain abnormalities.

3   Experiments

3.1   Materials

All images used in this work were posterior-anterior chest radiographs collected from a TB screening program among a high risk population. Radiographs were acquired with mobile digital thorax units (Delft Imaging Systems, the Netherlands) developed for cost-effective thorax examination and TB preventive screening. Images have a resolution of 2048×2048 and 12 bits data depth. Each image

was read by two radiologists, and a person whose radiograph was considered TB suspect by one of them or both was contacted to undergo further tests. For a subset of the cases, positive microbiological culture tests were available and a definite diagnosis of TB could be established.

We collected all TB suspect and TB proven cases between 2002 and 2005, and a similar amount of randomly selected normal radiographs, excluding radiographs of children. Before collection, radiographs were anonymized. Normal and TB suspect images were re-read by a third radiologist, who classified a part of the cases differently. Re-classified images were excluded from the study. Finally, our database contained 256 normal radiographs (223 males, 33 females, ages 18–70 yrs, median age 41), 178 TB suspect radiographs (155 males, 23 females, ages 16–101 yrs, median age 35), and 37 radiographs with microbiologically proven TB (30 males, 7 females, ages 16–43 yrs, median age 29).

3.2   Local Feature Extraction

For practical considerations, images were downsized to 1024×1024. Prior to feature extraction, lung fields were automatically segmented from the radiographs using multi-resolution pixel classification, with settings as given in [5]. In order to train this segmentation procedure, lung fields were segmented manually from 20 radiographs not used otherwise in this study.
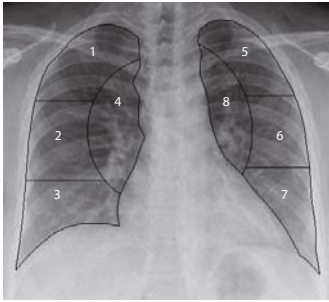
Next, local texture features were extracted from a large number of regions of interest (ROIs). At first, images were filtered with a multiscale filter bank of Gaussian derivatives, and subsequently central moments of histograms were calculated from each ROI in the original and the filtered images. The following parameters were chosen: Gaussian derivatives of orders $0, 1$ and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$ pixels; overlapping circular ROIs with a radius of 32 pixels placed on a grid with 8×8 pixel spacing inside the lung fields; and four central moments, namely, the mean, standard deviation, skewness and kurtosis. Before filtering, pixel values in the lung fields were mirrored outside the lungs symmetrically with respect to the lung borders in order to prevent contamination of extracted features due to strong filter responses at the lung border. Two position features were added that defined x and y coordinates of the ROI centers relative to the center of the mass of a lung field. In total, 126 features were extracted from each ROI, and the number of ROIs per image ranged from 1920 to 8680.

3.3   Lung Partitioning

In order to perform the MVDB classification on lung subdivisions, each lung field was automatically divided into 4 equal-sized regions (see Figure 2). The regions around hilum (regions 4 and 8) included lung pixels overlapping with a circle placed at the lungs' center of mass. The radius of the circle was separately chosen for the left and right lung, such that the overlap covered one quarter of the pixels of that lung. The rest of each lung field was horizontally divided into three equal-sized parts. The third radiologist assessed regions in all the TB suspect and TB proven images, and assigned a region to class 1 if a TB-related abnormality was present in the region, or to class 0 otherwise.

3.4   Classification

Training and test images were randomly selected from the available normal and TB suspect data, so that the training and test sets each contained 128 normal and 89 abnormal images. The second test set was formed from the same normal images as in the first test set and all 37 radiographs with proven TB. We randomly selected 10 normal and 10 abnormal radiographs to serve as prototype images. For region classification, the same division into training and test sets was used, but the random selection of prototypes was performed separately for each region, so that 10 normal and 10 abnormal regions were selected each time. Prototypes were always selected from the training images. Normal prototype regions were selected from normal training images only. Normal regions from abnormal images were excluded from the training set during region classifications.

The histograms of each local feature were obtained by a suitable binning of the range of feature values, either across the lung fields, or across a particular region. The range of possible values of each feature was estimated on prototypes and split into equal intervals - 128 for the lung fields, 64 for regions. A dissimilarity between two histograms was computed using χ^2 statistics as a dissimilarity measure:

Fig. 2. Division of the lung fields into eight regions

$$d_{\chi^2}(h, k) = \sum_i \frac{\left(h(i) - m(i)\right)^2}{m(i)}, \qquad (3)$$

where $h = \{h(i)\}$ and $k = \{k(i)\}$ are two corresponding histograms, i is a bin index, and $m(i) = \frac{h(i)+k(i)}{2}$. Dissimilarity representations were classified by the linear discriminant classifier. Classification was preceded by a principal component analysis (PCA) retaining 99% of variance to the dissimilarity representation, for the purpose of dimensionality reduction.

The MVDB method was compared with a straightforward approach where an image classification was composed of classification of each ROI and subsequent fusion of ROIs' posterior probabilities. In this approach, local feature vectors extracted from ROIs as described in Section 3.2 were used as input of the linear discriminant classifier preceded by the PCA. The division into training and test images was the same as for the MVDB experiments. ROIs from training images got the class labels of lung subdivisions they belonged to. An overall image decision was obtained by integrating all ROIs' posterior probabilities using the 95% percentile rule.

4   Results

The classification performance was estimated by means of the area under the receiver operating characteristic (ROC) curve, A_z [6]. A_z values for two test sets

Table 1. The performances of the MVDB classification and fusion, in terms of A_z

Test set	Lungs	Regions								Vote rule
		1	2	3	4	5	6	7	8	
Normal vs. suspect TB	**0.81**	0.79	0.71	0.85	0.66	0.82	0.81	0.72	0.77	**0.83**
Normal vs. proven TB	**0.70**	0.85	0.71	0.95	0.64	0.66	0.43	0.49	0.65	**0.74**

are presented in Table 1. The first column contains the results of the application of the MVDB classification to the whole lung fields. In the columns titled "1" to "8", A_z values for corresponding lung regions (see Figure 2) are listed. The final classification performance computed after combining global and local posterior probabilities by voting is given in the last column.

Combining global and local classification decisions slightly improves the overall classification performance compared to the results after applying the MVDB method to the whole lung fields only. To illustrate the gain of using the combination of local and global classifications, an example of a region with a slight diffuse abnormality is shown in Figure 3. This region was correctly classified as abnormal by the MVDB method applied locally (posterior probability $p_{c=1} = 0.89$), while an image containing this region was initially misclassified as normal by a global MVDB classifier ($p_{c=1} = 0.24$). After combing global and local results, the image received a probability of 0.89 of being abnormal.

The straightforward classification approach achieved $A_z = 0.77$ on the first test set and $A_z = 0.64$ on the second test set. This demonstrates the advantage of the MVDB method for classification of weakly labeled images.
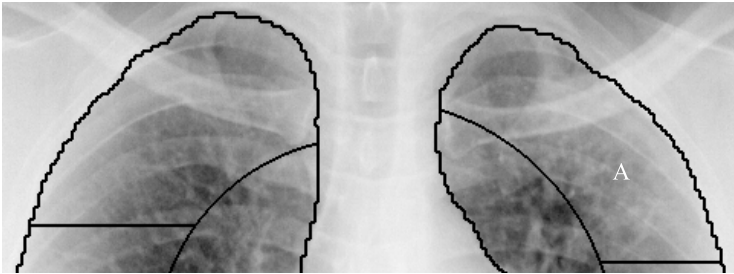
Fig. 3. An example of a correctly classified abnormal region (white "A" marks a proven TB lesion). An opposite region in the other lung is normal.

5   Discussion and Conclusions

The results presented in Table 1 demonstrate that the classification performance on the whole lung fields, as well as the performance of the combining scheme is considerably better on the first test set than on the second one. This observation can be explained in two ways. Firstly, the training set did not contain proven TB cases and so the test set with proven TB cases is expected to perform

worse. Adding proven TB cases to the training set is expected to improve the performance of the CAD system. Secondly, there were TB proven images with extremely subtle abnormalities in our collection. Such images are difficult for humans and CAD systems to classify.

The other observation is that the local performances vary greatly for both test sets, from 0.66 to 0.85 on the first set, and from 0.42 to 0.95 on the second set. Such a variation can happen due to low numbers of abnormal samples for some regions in a test set (e.g. the second test set contained only 3 abnormal regions "7" and only 3 abnormal regions "3"). Each misclassification then drastically influences an A_z value for such regions. For some regions, the number of abnormal samples in the training set was also limited, which in general negatively affected the MVDB classification performance on such regions. Future work will include the collection of a much larger data set which we expect to be beneficial for our combination scheme. In future, we should also investigate whether performing the selection of prototypes systematically can improve the results of our method, and how its performance is influenced by the number of prototypes.

In conclusion, we have shown that the multi-valued dissimilarity-based classification is a practical tool that enables a CAD system to deal with weakly labeled images. Combining global and local classification decisions has a potential to improve the overall classification performance. We have been the first to apply such a scheme to the automated detection of tuberculosis in a large database of digital chest radiographs.

References

1. World Health Organization: WHO Report 2008: Global Tuberculosis Control, Surveillance, Planning, Financing (2008)
2. van Ginneken, B., Katsuragawa, S., ter Haar Romeny, B.M., Doi, K., Viergever, M.A.: Automatic detection of abnormalities in chest radiographs using local texture analysis. IEEE Transactions on Medical Imaging 21(2), 139–149 (2002)
3. Pekalska, E.: Dissimilarity representations in pattern recognition. PhD thesis, Delft University, the Netherlands (2005)
4. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
5. van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis 10(1), 19–40 (2006)
6. Metz, C.: ROC methodology in radiologic imaging. Investigative Radiology 21(9), 720–733 (1986)