Multi-classifier semi-supervised classification of tuberculosis patterns on chest CT scans

E. M. van Rikxoort[1], M. Galperin-Aizenberg[1], J.G. Goldin[1], T.T.J.P. Kockelkorn[2], B. van Ginneken[2,3], M.S. Brown[1]

[1]Department of Radiological Sciences, University of California Los Angeles, United States
[2]Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[3]Diagnostic Image Analysis Group, Radboud University Nijmegen Medical Centre, The Netherlands

Abstract. Classification of different textures present in chest CT scans of patients with pulmonary tuberculosis (TB) is of crucial importance for the success of ongoing vaccine and drug testing trials. In this paper, a new multi-classifier semi-supervised method (MCSS) is proposed that is trained with a small set of labeled examples and improves classification performance by sampling interesting samples from unlabeled scans based on uncertainty among a pool of classifiers. The interesting samples are added to the small labeled set with a label assigned by 'expert' classifiers. MCSS is applied to 20 scans of patients with proven TB for which a reference standard was obtained by a consensus reading. Another set of 35 scans was used without manual labels. The performance of MCSS is compared to conventional supervised classification and two other semi-supervised methods and shown to outperform all other methods.

1 Introduction

Pulmonary tuberculosis (TB) is a major cause of morbidity and mortality world wide with 9.4 million new cases and 1.8 million deaths reported in 2008 [1]. Computed Tomography (CT) imaging is the most sensitive imaging technique for monitoring lung disease and can be used both for detecting and evaluating the progression of TB. On chest CT scans, TB presents as a wide variety of textural abnormalities. Quantifying the extent of TB is hard and time consuming even for expert radiologists. In addition, TB is most frequent in regions in the world where not many expert radiologists are available. Therefore, the development of computer aided diagnosis systems for detecting and quantifying TB is of crucial importance for the success of ongoing vaccine, drug testing, and screening programs. Some research toward detection and quantification of TB from chest radiographs has been performed, e.g. [2], however, no previous work on automatic quantification of TB from chest CT scans is available.

The large amount of data, combined with the difficulty of obtaining expert annotations, makes the problem of quantification of TB an excellent candidate for semi-supervised learning approaches. Semi-supervised learning is a popular technique in pattern recognition in which the performance of a classifier is improved by learning from unlabeled data, next to labeled data. There are a plethora of semi-supervised methods available, an overview can be found in [3]. Two common paradigms are self-training and co-training. In self-training, a classifier is first trained with a small amount of labeled data. The trained classifier is then applied to the unlabeled data and the samples for which the classifier is most confident about the label are added to the labeled set. This process is iterated several times. A problem of self-training is that it mainly enforces already known knowledge and errors in the classifier. Co-training [4] requires the feature set to be divided into two sets that are conditionally independent. Using these two feature sets, two classifiers are trained, and each classifier teaches the other with unlabeled examples of which it is sure of the label. The assumption underlying co-training that conditionally independent feature sets exist is a limiting factor for many applications. Therefore, several studies have been performed applying so-called multiview learning, in which multiple models are trained using the same set of labeled data, e.g. [5, 6].

In this paper, a variation of multiview learning is proposed in which an active learning based uncertainty sampling strategy to find interesting samples in the unlabeled data to be added to the labeled set is used. Instead of the human experts used in active learning, a set of three classifiers which can be regarded 'expert' classifiers are employed. The main contributions of this paper are the uncertainty based selection method of unlabeled samples for semi-supervised learning and the application of semi-supervised learning to a multi-class problem. A similar approach coined 'tri-training' was proposed for two-class problems in [6]. In tri-training, three classifiers are trained. If two classifiers agree on the label, the label is added to the training dataset of the third classifier not taking into account the label or confidence of the third classifier. A disadvantage of tri-training is that if the models in tri-training are not sufficiently different, the method degenerates to single-classifier self-training.

Multi-classifier semi-supervised classification (MCSS) is applied to the classification of TB patterns in 20 scans from 20 different patients enrolled in a vaccine-testing trial. For all 20 scans manual annotations were obtained by a consensus reading of two expert radiologists. Scans of 35 additional patients were used as unlabeled data. The results of MCSS are compared to conventional supervised classification, multi-classifier self-training, and multiview learning with majority voting in a cross-validation procedure.

2 Materials

In this paper, 55 scans from 55 different patients with smear positive TB were used. All scans are low dose CT (30mAs at 120 kV), reconstructed to 512×512 matrices with a pixel size of 0.7×0.7 mm and a slice spacing and slice thickness of 1 mm. As a preprocessing step, the lungs in all scans were automatically segmented [7]. Next, the lungs in all scans were automatically divided into small volumes of interest (VOIs) with roughly similar texture [8], which was shown to outperform square regions of interest. On average 4170 VOIs with an average volume of 1.2 ml per VOI were produced per scan. For both the manual labeling and automatic method these volumes will be classified instead of single voxels.

A team of expert radiologists predefined a set of five textures that can be found in chest CT scans of patients with TB: normal lung, consolidations, nodules (cavitated and non-cavitated), TB fibrosis, and (small) airway disease. Examples of each texture class are shown in Figure 1. It can be appreciated that especially in the classes nodules and airway disease there is some variation in the appearance of the lung parenchyma. For 20 out of the 55 scans, manual annotations were obtained by a consensus reading

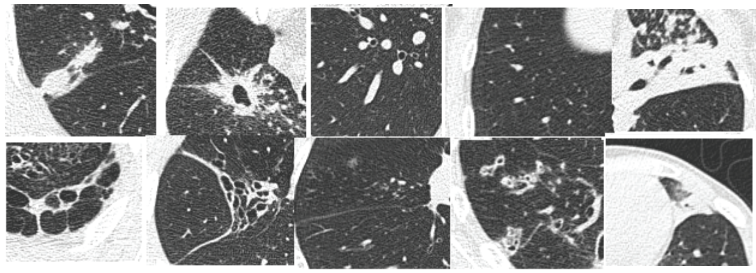

Fig. 1. Example patches of the different texture categories used. The first two images in the top row show examples of nodules (NOD), the second two images show normal lung regions (NL), the last image shows a consolidation (CONS). In the bottom row, the first two images show TB fibrosis (TBF), the second two images are examples of airway disease (AIR), an finally the last image is another example of a consolidation. It can be appreciated that even within the classes textures vary.

as follows: first, one expert radiologist annotated all 20 scans. Next, the second expert radiologist joined and together they went over all annotated regions. In case of doubt a consensus reading was performed. Since annotating is very time consuming and not all patterns occur in each scan, the radiologists were instructed to freely annotate in each scan classical examples of the available textures. Counters of the lung volume annotated were shown to the radiologist. In total, the first radiologist spent 8.5 hours annotating the scans. Consensus reading took another 3 hours. In total 2148 volumes were annotated after consensus reading. The division of the labeled VOIs over the different texture classes was: normal lung 423, consolidations 19, nodules 362, TB fibrosis 304, and airway disease 1040.

3 Methods

All supervised methods consist of a training phase, in which the classifier is trained, and a test phase in which the trained classifier is applied to test data. The difference between the different methods described in this section is the training phases, the outputs of all systems is a trained classifier, which is applied to test data to label each VOI in the test scans.

This section consists of four parts. First, the features and classifiers used for all methods are provided. In the second part, a conventional supervised texture classification system is described followed by a description of the proposed multi-classifier semi-supervised classification method in the third part. Finally, the semi-supervised methods implemented for comparison are briefly introduced.

Features and classifiers For each VOI, the first four statistical moments (kurtosis, skew, mean, standard deviation) of a set of 14 features on four scales were used, resulting in 224 features in total. The image features used were the output of Gaussian filters up to and including second order derivatives (L, L_x, L_y, L_z, L_{xx}, L_{xy}, L_{xz}, L_{yy}, L_{yz}, L_{zz}),the gradient (L_i), and the eigenvalues of the Hessian matrix ($|\lambda_0| \geq |\lambda_1| \geq |\lambda_2|$). All features were calculated on scales 1, 2, 4, and 8. Classification

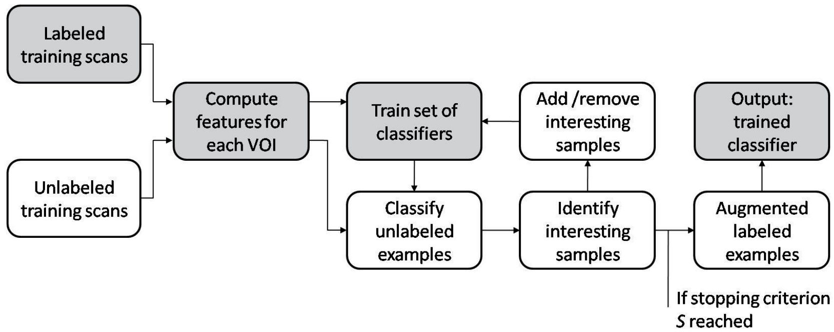

Fig. 2. Flowchart of multi-classifier semi-supervised (MCSS) segmentation. Steps indicated in gray are the steps performed during conventional supervised classification.

was preceded by a principal component analysis (PCA) retaining 95% of variance for the purpose of dimensionality reduction.

For training purposes, the initial labeled set was randomly sub sampled to contain equal class sizes for all 5 classes. However, since for one class, consolidations, the number of samples is substantially lower than for the others, this class is not taken into account during subsampling. As a result, all classes contain 304 labeled examples except consolidations, which contains 19 labeled examples. All 2148 labeled VOIs will be used for evaluation in a leave-one-patient-out cross validation procedure.

For conventional supervised texture classification a support vector machine (SVM) classifier with radial basis kernel functions was used as a classifier. The settings of the SVM were determined on the initial labeled set using cross validation [9]. For MCSS and other semi-supervised methods a linear discriminant classifier (LDC) and k-NN classifier (KNN) with k equal to 7 were added to the pool of classifiers. The final classifier after all semi-supervised methods is the SVM trained with the extended labeled set.

Conventional supervised texture classification (CVS) In the training phase of CVS, the set of features is computed for all VOIs which have been labeled. Based on the features and the known output labels, the SVM classifier is trained to be able to assign a label to previously unseen VOIs. In the flowchart in Figure 2 the steps to train a classifier in a conventional supervised texture classification system are indicated in gray.

For a test scan, the feature vector is calculated for each VOI and the trained SVM classifier assigns a label to each feature vector.

Multi-classifier semi-supervised texture classification (MCSS) The goal of semi-supervised classification is to extend the initial labeled set using unlabeled examples to increase classifier performance. A flowchart of the proposed semi-supervised method is provided in Figure 2. Globally, the procedure of MCSS is as follows: given a set of labeled VOIs, L, and a set of unlabeled VOIs, U, a pool of n classifiers $C_i,\ i = 1, \ldots, n$ is trained using L. The trained classifiers C_i are applied to all samples s_u of U. Inspecting the posterior probabilities p of each C_i for each s_u, 'interesting' s_u are identified that should be removed from U and added to L with the label assigned by the pool of classifiers. This process is iterated until a stopping criterion S is reached.

When S is reached, the extended labeled set L is used to train a final classifier. The key point of this scheme is the identification of 'interesting' samples in the unlabeled data.

We propose to use a paradigm that is used in active learning to identify interesting samples: if there is uncertainty about the label of a sample, this is an informative sample and an expert opinion should be obtained. In MCSS we view the set of classifiers used as individual experts. Uncertainty is defined in two ways: if one of the classifiers is unsure about the label to assign but the other classifiers agree on the label with high confidence, the sample is added to the labeled set with the label of the agreeing classifiers. Or secondly, uncertainty about the label exists if two classifiers agree with a high confidence on the label of a sample but the third classifier is confident about another label. In this case the sample is added to the labeled set with the label of the agreeing classifiers. The labeled example is added to the training dataset of all classifiers, which leads to a combination of majority voting (all but one classifiers agree with high confidence) and uncertainty sampling (only if the remaining classifier is unsure or disagrees on the label). The rationale behind this approach is twofold. First, samples for which uncertainty exist are the most informative since they change the classifier as opposed to samples for which all classifiers agree. And second, the high agreement between the two 'expert' classifiers makes it more likely that they made the right decision. MCSS is implemented by setting three parameters: the posterior probability p at which a classifier is confident, $p > p_c$, and the posterior probabilities between which a classifier is unsure of its label, $p_{ul} < p < p_{uh}$. For classification of TB textures, p_c was set to 0.6 and p_{ul} and p_{uh} were set to 0.2 and 0.5, respectively.

Once all unlabeled samples have been processed, the classifiers are retrained with the extended labeled set and the process is iterated. Due to the unbalanced appearance of the different structures in the data, a pruning step is performed in each iteration of MCSS; the classes are pruned to be of equal size, the size being the size of the one but smallest class. Since all classifiers used are provided new examples, they are refined in each iteration and therefore different results will be obtained after each iteration. It is important to note that the classifiers used in this scheme should be diverse since if their labeling of unlabeled samples is identical, no interesting features can be identified. The final parameter to be set is a stopping criterion S for iterating this process. In this paper, S was set to no more interesting samples being identified for at least 2 of the classes. As a final classifier, any of the classifiers used can be trained or a combination of them can be used. In this paper the SVM classifier trained with the final extended labeled set is used as a final classifier.

It is obvious that if the class label assigned by the agreeing classifiers is correct, the training data is augmented with a sample with a valid label. Otherwise, a noisy label will be added to the training dataset. In [6] it is shown that if the amount of newly added examples is sufficient, the increase in classification noise can be compensated for.

Semi-supervised methods implemented for comparison For comparison, two other semi-supervised classification methods were implemented: multi-classifier self training (MCST) in which a label is added when all classifiers agree with high confidence, and multiview majority voting (MMV), in which a label is added if at least two classifiers agree with high confidence. For MCST and MMV the same p_c as for MCSS was used.

4 Experiments and Results

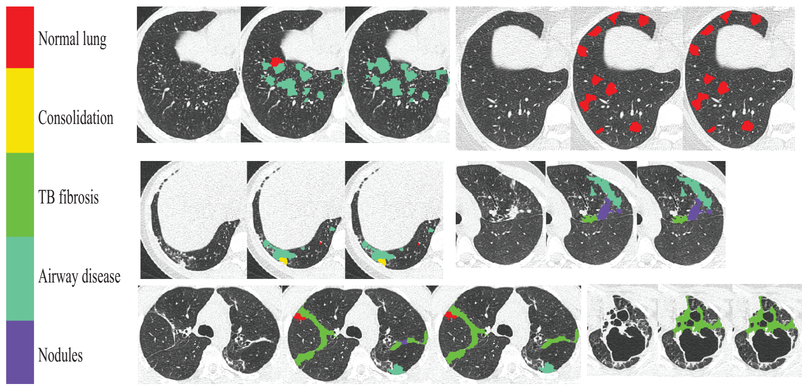

Fig. 3. Example output of MCSS and the ground truth for several slices of different scans. For each group of 3 slices, the left slice is the original slice, the middle slice is the output of MCSS, and the last slice shows the ground truth. Two wrong assignments can be seen in these examples: the top left example shows a VOI that was labeled as normal lung by MCSS and as airway disease by the observers. The left images in the bottom row show an example of fibrosis mislabeled as lesion, which is the most common mistake of MCSS.

All experiments were performed in a leave-one-patient out cross validation procedure. On average, MCSS performed 5 iterations, adding 6024 samples to the labeled set in total. 1477 samples were added with the label nodules, 50 samples were added to the consolidations class, and 1499 labels were added to the other three classes. The equal number for the last three classes is due to the pruning that is performed after each iteration. To show the validity of the classifiers used as experts during semi-supervised learning, we calculated their accuracy on the initial labeled set for those labels for which $p > p_c$. The LDC classifies 93% of the samples with $p > p_c$ with an accuracy of 0.83. For the KNN classifier 66% of the samples has $p > p_c$ with an overall accuracy of 0.81, the SVM classifier classifies 88% of the samples with $p > p_c$ with an accuracy of 0.85.

Figure 3 shows for several scans an original slice, the result of MCSS, and the manual labeling. Table 1 provides the confusion matrices for CVS, MCSS, MCST, and MMV. The accuracies for the different classes are provided next to the confusion matrix, both in total and per class. Next to overall improvement in performance, it can be seen that especially classes with a low accuracy (consolidations and TB fibrosis) in the CVS get boosted when applying MCSS. In addition, MCSS performs better than the other semi-supervised methods, not only in total but also for every class separately; there is no class for which one of the other methods performs better.

Table 1. Confusion matrices and accuracies for CVS, MCSS, MCST, and MMV. The rows depict the ground truth, columns the output of the automatic methods.

		CVS					acc	MCSS					acc
		N	**C**	**N**	**F**	**A**	0.816	**N**	**C**	**N**	**F**	**A**	0.863
	NL	405	0	0	4	14	0.957	418	0	0	0	5	0.988
	CONS	0	11	7	0	1	0.579	0	17	2	0	0	0.895
GT	**NOD**	5	2	270	75	10	0.746	3	0	302	54	3	0.834
	TBF	8	0	87	185	24	0.609	2	0	51	235	16	0.773
	AIR	74	0	32	51	883	0.849	70	0	42	45	883	0.849
		MCST					**acc**	**MMV**					**acc**
		N	**C**	**N**	**F**	**A**	0.831	**N**	**C**	**N**	**F**	**A**	0.837
	NL	409	0	0	1	13	0.967	416	0	0	1	6	0.983
	CONS	0	15	3	0	1	0.789	0	17	2	0	0	0.895
GT	**NOD**	6	0	282	65	9	0.779	4	0	275	73	10	0.760
	TBF	8	0	59	214	23	0.704	11	0	58	214	21	0.701
	AIR	99	0	29	47	865	0.831	88	0	30	44	878	0.844

5 Conclusion & Discussion

This paper presents a multi-classifier semi-supervised approach (MCSS) for classification of TB textures on chest CT scans. The problem of TB classification is highly appropriate for a semi-supervised approach due to the difficulty of obtaining manual annotations. The main contribution of this paper is the selection of interesting unlabeled samples to add to the labeled set based on uncertainty sampling. The proposed method performs well for the task of classification of TB textures and outperforms conventional supervised classification as well as other well-known semi-supervised methods. The proposed method of identifying interesting samples from unlabeled data is especially fit for multi-class problems since classifiers are more often unsure in these cases.

A limitation of this study is the relatively small number of scans that were manually annotated. In addition, due to the difficulty of the task, the the observers only indicated classical examples of the different textures. This potentially makes the classification task easier and boosts the performance of the automatic classification. To improve the manual annotations for future work, an active learning step will be introduced in which observers are asked to annotate VOIs for which the classifier is unsure.

Several parameters have to be set in MCSS. The influence of the setting of these parameters has not been studied in this paper. In general, the higher p_c, the less samples will be added to the labeled data but the more confident the classifiers are about the assigned label. For the application of TB classification setting p_c above 0.90 leads to only sampling normal VOIs from the unlabeled data. The setting of p_{ul} and p_{uh} in this paper were based on the fact that a five class classification task was performed; a posterior probability between 0.2 and 0.5 in a five class problem indicates that the posterior probabilities are relatively spread over at least 3 classes.

For any semi-supervised method it is important that the samples added to the labeled set have correct labels. Since it is unavoidable that also samples with incorrect labels are added to the labeled set, it is important to have a large pool of unlabeled data. In this paper a set of 35 unlabeled scans was used. Due to the variation in appearance of the abnormalities and the difference in their prior probability (e.g. consolidations vs.

airway disease) future work includes extending the set of unlabeled scans used during MCSS.

To conclude, a multi-classifier semi-supervised classification method was presented that was applied to the classification of TB texture patterns. The proposed method was shown to be able to increase classification performance compared to conventional supervised classification by adding unlabeled samples to the labeled dataset and outperform other semi-supervised methods implemented for comparison.

References

1. World Health Organization: WHO Report 2009: Global tuberculosis control, Epidemiology, Strategy, financing (2009)
2. Arzhaeva, Y., Hogeweg, L., de Jong, P., Viergever, M., van Ginneken, B.: Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis. In: Medical Image Computing and Computer-Assisted Intervention. Volume 5762 of Lecture Notes in Computer Science. (2009) 724–731
3. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
4. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. (1998) 92–100
5. Zhou, Y., Goldman, S.: Democratic co-learning. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. (2004) 594–602
6. Zhou, Z.H., Li-Yueh, M.: Tri-Training: Explointing unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering **17**(11) (2005) 1529–1541
7. van Rikxoort, E.M., de Hoop, B., Viergever, M.A., Prokop, M., van Ginneken, B.: Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Medical Physics **36**(7) (2009) 2934–2947
8. Kockelkorn, T.T.J.P., de Jong, P.A., Gietema, H.A., Grutters, J.C., Prokop, M., van Ginneken, B.: Interactive annotation of textures in thoracic CT scans. In: SPIE Medical Imaging. Volume 7624. (2010) 76240X1–76240X8
9. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.