Semi-supervised based learning for Idiopathic Interstitial Pneumonia on High Resolution CT images

SHOUNO HAYARU[1,a)] KIDO SHOJI[2,b)]

Abstract: In the classification task, the number of labeled samples is one of the important factor for accuracy, however, gathering such data is hard work since it requires diagnosing task in the field of medical engineering, In order to overcome this problem, we introduce a semi-supervised learning (SSL) classifier for computer aided diagnosis (CAD) for idiopathic interstitial pneumonias (IIPs). The semi-supervised learning requires a lot of unlabeled training data, which does not require diagnosing cost, as well as labeled data. In this study, we show the low performance classifier, which has only chance level classification performance, would be improved to achieve around 90% accuracy performance by SSL. We also propose a pre-processing method of gray-scale transformation for appropriate application to the SSL. Without proper gray-scale transformation, the SSL might cause decreasing performance however, we find our pre-processing procedure make increasing the performance in almost all the cases.

1. Introduction

In the medical diagnosis, the classification task is important for the diagnosis quality. For classifying and detecting the idiopathic interstitial pneumonias (IIPs), high-resolution computed tomography (HRCT) image is regarded to be effective since IIPs affected part looks diffused in the lung [1][2][3][4][5]. Unfortunately, determining the border of the site is difficult work, because the IIPs on HRCT images show a lot of varieties in the meaning of texture patterns. The quality of diagnosis is influenced by the ability of physician, and improving the quality is desired for proper treatment of IIPs. In order to decrease the burden of physicians, development of the computer aimed diagnosis (CAD) system is desired for objective diagnosis in these decades [1][6][5]. The CAD systems are designed to provide a classification function for second opinion using machine learning techniques.

In the field of machine learning, the supervised learning is usually used for such classification task. and it requires pairs of input patterns and its corresponding labels for its learning. For improving the classification performance of such supervised learning system, a lot of labeled learning data is required, however, the obtaining cost of such data is expensive since it requires physicians diagnoses to get the proper labels. On the contrary, the cost for obtaining unlabeled data, which does not require physicians diagnosis, is lower than that of the labeled. The semi-supervised learning (SSL) uses massive unlabeled learning data as well as the labeled in order to improve the performance of classification accuracy [7].

[1] Graduate School of Informatics and Engineering, University of Electro-Communications, Chofugaoka 1-5-1, Chofu, 182–8585, Japan
[2] Graduate School of Medicine, Yamaguchi University, Tokiwadai 2-16-1, Ube, 755–8611, Japan
[a)] shouno@uec.ac.jp
[b)] kido.ai@yamaguchi-u.ac.jp

The IIPs sites in the HRCT images are usually diffused in the lung, so that, we can obtain these unlabeled data easily by slightly shifting of the labeled region of interests (ROIs). Our purpose is to improve the accuracy performance of the CAD system for IIPs classification using the SSL by use of such unlabeled data. In this study, we try to develop and evaluate a classifying engine for IIPs in the CAD system.

2. Method

2.1 Semi-Supervised Learning system

In this study, we denote the input feature as a vector $\boldsymbol{x}$, and its desired label as t. The supervised learning data are denoted as pairs of the features and labels, $\{\boldsymbol{x}_n, t_n\}$, where n means the index. On the contrary, we use $\boldsymbol{x}^u$ as the unlabeled input feature, and denote the unlabeled data as $\{\boldsymbol{x}^u_m\}$ where m means the index.

The fig.1 shows the schematic diagram of a simple SSL, which is called "self-training" architecture proposed by Yarowski [8][9]. In the SSL, at first, a supervised classifier is trained by only use of labeled data $\{\boldsymbol{x}_n, t_n\}$. In the next step, the supervised classifier predict labels for the unlabeled features $\{\boldsymbol{x}^u_m\}$ with beliefs, which mean the confidences for the labels of the classifier. Thus, the unlabeled data can be regarded as the labeled $\{\boldsymbol{x}^u_m, t^u_m\}$, where t^u_m is the label by the supervised classifier. After that, we drop off the

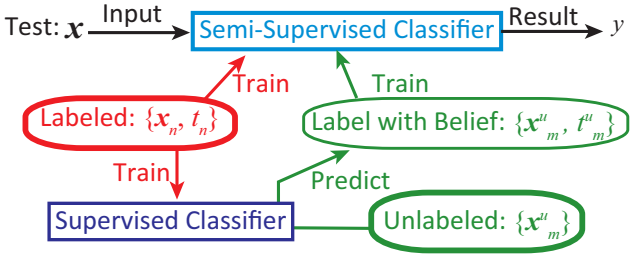

Fig. 1 Schematic diagram of the SSL architecture (Yarowski, 1995)[8], [9]

low belief data by thresholding. The SSL classifier is trained by both the labeled data for supervised learning and classified unlabeled data with high beliefs.

This simple algorithm might be applied to the previous works for IIPs classification[5]. However, it is hard to evaluate of the genuine ability of the SSL by use of the complex learning system, so that, we apply a simple naive naive Bayes classifier in this study. Assuming the predicting label as y for the input feature $\boldsymbol{x}$, the Bayes classifier calculate posterior probability of $P(y \mid \boldsymbol{x})$ by use of the Bayes' theorem, that is, we can denote the posterior probability as $P(y \mid \boldsymbol{x}) \propto P(\boldsymbol{x} \mid y)P(y)$ where $P(\boldsymbol{x} \mid y)$ and $P(y)$ mean the likelihood and prior probability respectively. The prior $P(y)$ is defined as $P(y = k) = n_k/N$ where k means the class label and n_k means the number of labeled images belonging to the class k in the training set. N means the total number of the labeled images of training set $N = \sum_k n_k$. The likelihood function $P(\boldsymbol{x} \mid y = k)$ is derived the multi-dimensional Gauss distribution $P(\boldsymbol{x} \mid y = k) \sim \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{m}_k, \Sigma_k)$ where $\boldsymbol{m}_k$ and Σ_k means average of feature vectors in the class k and corresponding covariance matrix respectively.

In the SSL, the Bayes posterior probability works as the beliefs for the unlabeled input feature $\boldsymbol{x}^u_m$. The class label for these unlabeled inputs are given by the supervised classifier with maximization of the posterior probability: $t^u_m = \mathrm{argmax}_k P(t^u_m = k \mid \boldsymbol{x}^u_m)$. In this maximization process, we obtain the probability value $P(t^u_m = k \mid \boldsymbol{x}^u_m)$ which means the confidence for the classification label, so that, we treat this value as the belief for the label $t_m = k$.

3. Experiments

3.1 Materials

In order to construct the SSL classifier, we prepare 360 labeled images and 3600 unlabeled images. In the labeled images, the number of each class are following: Consolidation(CON):38, Ground-Grass-Opacity (GGO):76, Honeycomb(HCM):49, Reticular(RET):37, Emphysema(EMP):54, Nodular(NOD):48, and Normal(NOR):58 cases. We assume the 32×32 [pixels] ROIs, and each ROI is segmented under the direction of a physician, and diagnosed by 3 physicians.

The acquisition parameters of those HRCT images are as follows: Toshiba "Aquilion 16" is used for imaging device, each slice image consists of 512×512 pixels, and pixel size corresponds to $0.546 \sim 0.826$ [mm], slice thickness are 1 [mm]. The number of patients is 69 males and 42 females with age 66.3 ± 13.4. The number of normal donor is 4 males and 2 females with age 44.3 ± 10.3. The origin of these image data is provided Tokushima University Hospital. Fig.2 shows a typical image example of each disease in HRCT image. The left shows an overview of the axial HRCT images of lungs including lesion, and the right shows segmented images of typical examples of lesion from the left image collections. The consolidation (CON) and ground-grass opacity (GGO) patterns are often appeared with the cryptogenic organizing pneumonia diseases (COPD). The GGO pattern is also often appeared in the non-specific interstitial pneumonia (NSIP). The reticular (RET) pattern which sometimes includes GGO patterns is also appeared in the NSIP. The honeycomb (HCM) pattern has more rough mesh structure rather than that of the crazy-paving, and it appeared in idiopathic pulmonary fibrosis (IPF) or usual interstitial pneumonia (UIP).

3.2 Labeled and Unlabeled dataset

From the 360 labeled ROI images, we define the labeled dataset as followings. The labeled ROI images have been carried out gray-level transformation in order to diagnose by physician. The raw HRCT image pixel value, which is counted with Hounsfield unit (H.U.), is adjusted to describe the physical matters, and its resolution takes 4096 grades in the range of $[-1024, 3071]$ [H.U.]. For example, air takes -1024 [H.U.], water takes 0 [H.U.], and over over 500 [H.U] shows bone typically. In order to diagnose the lung, which mainly occupied with air, the full resolution of pixel value is too much information for diagnose, so that gray-level transformation is applied as pre-process. The gray-level transformation for the raw HRCT image pixel value I is described as piece-wise linear transformation:

$$q = \begin{cases} 0 & I < \mathrm{WL} - \frac{\mathrm{WW}}{2} \\ 255 & I > \mathrm{WL} + \frac{\mathrm{WW}}{2} \\ \frac{255}{\mathrm{WW}}\left(I - \left(\mathrm{WL} - \frac{WW}{2}\right)\right) & \text{else} \end{cases}, \qquad (1)$$

where q means the 256 grades gray image value. Thus, the gray-level transformation is controlled by the window-level (WL) and the window-width (WW) parameters. Typically, these parameters are defined for diagnosing part such like lung, and sometimes adjusted by the physician manually. The labeled data has been processed by $\mathrm{WL} = -600$[H.U.] and $\mathrm{WW} = 1500$[H.U] in order to adjust the pixel values of the ROI images in $[0, 255]$ range.

The labeled dataset named as L is randomly selected from each class k ($k = 1 \cdots 7$, which means the class label) evenly, and we denote the total number of the labeled dataset L as N. This labeled dataset L is used for the supervised learning of the Bayes classifier.

We prepare 3 unlabeled datasets as follows. The first dataset U_1 is simply use of the rest of labeled data, which consists of $360 - N$ images. The dataset U_1 is used for control dataset against other unlabeled. The other unlabeled dataset U_2 and U_3 come from the raw HRCT images. These unlabeled candidates are gathered from the surrounds of labeled ROIs site, and the total number of collected unlabeled image data becomes $3,600$. In order to use these images for training data, we should carry out the gray level transformation 1 as a preprocessing. In our gray level transformation, we assume the transformation parameters WW and WL are not given since these parameters, which are adjusted manually, are not rigid even in the labeled data. Thus, we should infer these parameters, and details are followings. The pixel of the raw HRCT image has 16-bit depth in this case and the range of unlabeled data pixels values are in $[-1152, 7281]$.

The parameters WL and WW for dataset U_2 are defined by these averaged histograms. We optimize the parameters WW and WL as to maximize the similarity between the labeled and transformed unlabeled histograms. We denote the average histogram of labeled as q_L, which can be regarded as a probability distribution. We also describe the transformed unlabeled histogram as the function of the parameters WW and WL, that is, $q_U(\mathrm{WW}, WL)$. Then, we introduce Kullback-Leibler (KL) divergence as a similarity measure between the gray level histograms

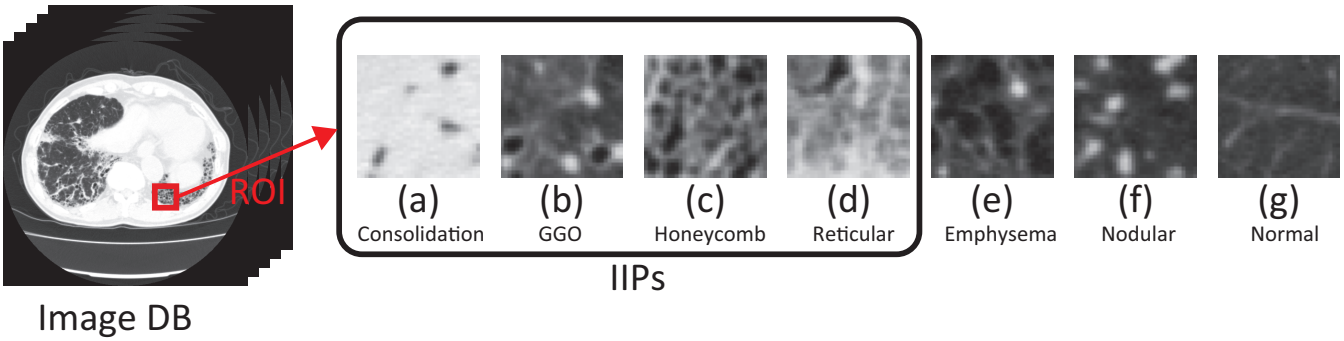

Fig. 2 Typical HRCT images of diffuse lung diseases: The top row shows each overview, and bottom shows magnified part (ROI) of each lesion. From (a) to (g) represents "Consolidation", "GGO", "Honeycomb", "Reticular", "Nodular" "Emphysema", and "Normal" image respectively.

q_L and $q_U(\mathrm{WW}, \mathrm{WL})$:

$$\mathrm{KL}(q_L \mid q_U; \mathrm{WW}, \mathrm{WL}) = \sum q_L \ln \frac{q_L}{q_U(\mathrm{WW}, \mathrm{WL})}. \qquad (2)$$

We adopt WW and WL as the minimization values of the $\mathrm{KL}(q_L \mid q_U)$. From the strategy, we optimize eq.2 and obtain the parameters WW = 1234[H.U.] and WL = −434 [H.U.] for the unlabeled data.

For comparison, we prepare the other unlabeled dataset name U_3 whose WW and WL are chosen as a typical values to observe lung-area, that is WW = 1500[H.U.] and WL = −550 [H.U.].

3.3 Feature Extraction and Selection

We introduce a texture analysis proposed by Sugata *et al.* for feature extraction [1][10]. From the input HRCT ROI image, we calculate gray-level histogram, gray-level difference statistics, the co-occurrence matrix, run length matrix, and Fourier power spectrum, at first. After that, from these 5 quantities, we derive 39 texture statistics as the candidates for features[10]. Using whole statistics candidates as the input features for classifier might cause the decreasing the performance because of "curse of dimensionality". Thus, we would select the 4 input features as the input for the classifier. These feature are determined experimentally.

3.4 Evaluation Method

In order to evaluate performance, we adopt leave-one-out cross-validation (LOOCV) [11][12]. In the LOOCV method, we choose a ROI image from the labeled dataset L and use other $N-1$ images for supervised learning. After supervised learning, M unlabeled images from the unlabeled dataset U_1, U_2, or U_3 is used for the self-training method. The preserved ROI image is used for evaluating the classification performance. This evaluation process is applied alternate to the whole labeled dataset L, and finally the average accuracy is used for the performance measure.

For the performance evaluation, we carry out the SSL method as follows:

(1) We trained Bayes classifier with supervised learning that is we apply only the labeled data. For LOOCV method, we pulled out a pair of training datum from labeled dataset. Then, we calculate the mean vector $\boldsymbol{m}_k$ and covariance matrix Σ_k for the pulled dataset.
(2) We predict the class label for the unlabeled dataset with confident that comes from the posterior probability. The largest posterior class is to become the predicted class.
(3) By thresholding, we drop out the low confident unlabeled data. We adopt the threshold value as 0.80 in this experiment.
(4) From the rest of the unlabeled data, which are high confident unlabeled data, we select M images randomly.
(5) We train the Bayes classifier with both labeled and the M predicted data again
(6) We evaluate the classifier accuracy by the selected labeled data pair in the procedure 1 (LOOCV method).

4. Results

We compare the accuracy performances among the added number of unlabeled data from U_1, U_2, and U_3 whose differences are gray-level transformation parameters of WW and WL. The dataset U_1 has identical statistical property since it comes from labeled data. The dataset U_2 might have similar property to the labeled data in the meaning of the KL-divergence. The dataset U_3 might have the most different property to the labeled data, however, it is typical parameters for observing lung areas. Fig.3(a) shows the accuracy performance against the added number of the unlabeled images M. The horizontal axis shows the number of the unlabeled images M in log-scale. The vertical one shows the accuracy performance. The size of labeled dataset L is $N = 35$. Adding small number of unlabeled data ($M \sim 100$) increase the accuracy performance in the every unlabeled dataset U_1, U_2, and U_3. The number of dataset U_1 is $360 - N$, so that, the curve ends in the value with high accuracy performance. The curve using the U_2 saturate around $M \sim 1000$ with also high accuracy performance. However, the curve using the dataset U_3 saturates in the low accuracy performance, while U_1 and U_2 increase the accuracy performance. In this case, the final accuracy performances for U_1, U_2, and U_3 datasets are 96.6%, 95.8%, and 61.1% respectively.

Fig.3(b) shows also the accuracy performance under the larger labeled dataset L that have $N = 70$. In this case the initial accuracy performance, which comes from the supervised classifier, is higher than the previous result, that is, the accuracy performance is 84.5% correct. We can see the similar tendency to the previous result in the meaning of the performance improvement, while the curve using U_3 decrease the its performance. In this case, the

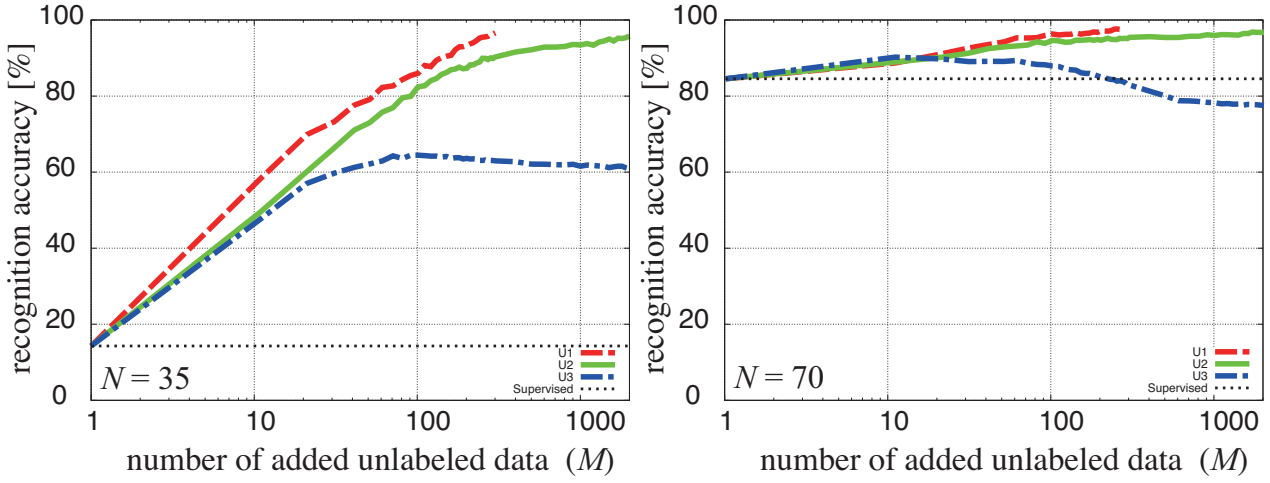

Fig. 3 Accuracy performance for several unlabeled dataset U_1, U_2, and U_3. The horizontal axis shows the number of added unlabeled images denoted as M. The vertical shows the accuracy performance. (a): The left shows the result with the supervised classifier that is trained by $N = 35$ samples. (b): The right shows the result for the classifier trained by $N = 70$ labeled samples.

final accuracy performances for U_1, U_2, U_3 dataset are 97.5%, 96.8%, and 77.5% respectively.

From these results, statistical similarity between the labeled image and the unlabeled images is important factor to the accuracy performance.

5. Conclusion & Discussion

We investigate classification performance of the SSL for the classification task of the IIPs. We can confirm increasing of the accuracy performance in several cases while the self training method is a simple method in the SSL.

We found several important factors for the IIPs classification by the SSL. One is the statistical quality of the features, that is, the gray-scale histograms of unlabeled images should have similar property to that of the labeled images used in the supervised learning. The unlabeled dataset U_2 optimized for the minimization of the KL-divergence between gray-level histograms of labeled and unlabeled. This result suggests that unlabeled data that come from another HRCT device might be available when we carry out appropriate pre-processing.

Moreover, we evaluate several trials for the another labeled dataset L, and stable improvement is confirmed by the SSL. When the initial supervised learning make good accuracy performance, the SSL improvement does not work well, however, it does not cause cause adverse affect in this investigation. Thus, we can consider the SSL is a good framework to improve classification ability for our task.

Acknowledgment

We thank Professor Junji Ueno, Tokushima University. He provided several advice for this study as well as a set of high resolution HRCT image of IIPs. This work is supported by Grant-in-Aids for Scientific Research (C) 25330285, and Innovative Areas 26120515, MEXT, Japan.

References

[1] Uppaluri, R., Heitmman, E., Sonka, M., Hartley, P., Hunninghake, G. and Mclennan, G.: Computer Recognition of Regional Lung Disease Patterns., *American Journal of Respiratory and Critical Care Medicine*, Vol. 160, No. 2, pp. 648–654 (1999).

[2] Kauczor, H. U., Heitmann, K., Heussel, C. P., Marwede, D., Uthmann, T. and Thelen, M.: Automatic detection and quantification of ground-glass opacities on high-resolution CT using multiple neural networks: comparison with a density mask, *AJR Am J Roentgenol*, Vol. 175, No. 5, pp. 1329–1334 (2000).

[3] Webb, W., Müller, N. L. and Naidich, D.: *High Resolution CT of the Lung, 4th edn.*, Lippincott Williams & Wilkins, Baltimore (2008).

[4] : American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001, *Am. J. Respir. Crit. Care Med.*, Vol. 165, No. 2, pp. 277–304 (2002).

[5] Xu, R., Hirano, Y., Tachibana, R. and Kido, S.: Classification of diffuse lung disease patterns on high-resolution computed tomography by a bag of words approach, *MICCAI*, Vol. 14, Springer-Verlag Berlin Heidelberg, pp. 183–190 (2011).

[6] Sluimer, I., Schilham, A., Prokop, M. and Ginneken, B.: Computer Analysis of Computed Tomography Scans of the Lung: A Survey., *IEEE Transactions on Medical Imaging*, Vol. 25, No. 4, pp. 385–405 (2006).

[7] Xiaojin, Z.: Semi-Supervised Learning Literature Survey, Technical report, Computer Sciences, University of Wisconsin-Madison (2005).

[8] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 189–196 (online), DOI: 10.3115/981658.981684 (1995).

[9] Haffari, G. and Sarkar, A.: Analysis of Semi-Supervised Learning with the Yarowsky Algorithm, *UAI* (Parr, R. and van der Gaag, L. C., eds.), AUAI Press, pp. 159–166 (2007).

[10] Sugata, Y., Kido, S. and Shouno, H.: Comparison of two-dimensional with three-dimensional analyses for diffuse lung diseases from thoracic CT images, *Medical Imaging and Information Sciences*, Vol. 25, No. 3, pp. 43–47 (online), available from ⟨http://ci.nii.ac.jp/naid/130000097652/en/⟩ (2008).

[11] Stone, M.: Cross-validation: A review., *Math.Operations.Stat.Ser.Stat*, Vol. 9, No. 1, pp. 127–139 (1978).

[12] Bishop, C. M.: *Pattern Recogition and Machine Learning*, Springer (2006).