International Journal of Soft Computing 4 (5): 173-184, 2009
ISSN: 1816-9503

Participatory Learning Based Tri-Training Algorithm for Computer Aided Diagnosis

C. Deng and M.Z. Guo
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Abstract: Tri-training is a promising semi-supervised learning approach for Computer Aided Diagnosis (CAD) systems. It aims at enhancing the performance of the hypothesis that is learned on only a small amount of expert-diagnosed samples by utilizing the large amount of undiagnosed samples through co-training process. However, mislabeling the unlabeled samples in the co-training process is inevitable and harms the performance improvement of the hypothesis. In this study, we extend the co-training process by a participatory learning cognition paradigm and propose a new tri-training algorithm named PL-Tri-training. In detail, the acceptance unit of participatory learning is instantiated as a data editing operation and the critic unit of participatory learning is designed as an adaptive arousal strategy for the data editing. In the co-training process of PL-Tri-training, the acceptance unit utilizes data editing to identify and remove the mislabeled data, as well as the critic unit exploits arousal strategy to inhibit the invalid activation of data editing. Experiments on three benchmark medical data sets verify the effectiveness of the proposed algorithm. A successful application to the pulmonary nodules detection in chest CT images shows that PL-Tri-training can more effectively exploit the undiagnosed samples to improve the diagnosis performance than Tri-training and AC-Tri-training, which extends the co-training process only with the acceptance unit of participatory learning.

Key words: Participatory learning, tri-training, semi-supervised machine learning, co-training, computer aided medical diagnosis, small pulmonary nodules detection

INTRODUCTION

Many machine learning techniques have been successfully applied to Computer-Aided Diagnosis (CAD) systems (Anagnostopoulos and Maglogiannis, 2006). To make the CAD systems perform well, these methods usually require a large amount of expert-diagnosed examples. These examples can be easily collected by routine medical examinations. However, making a diagnosis for such a large amount of cases one by one places a heavy burden on medical experts. For instance, to construct a CAD system for lung cancer diagnosis, radiologists have to label every focus in a huge amount of easily obtained high-resolution chest CT images. This task is usually quite time consuming and tedious. Therefore, the semi-supervised learning approaches become an attractive topic in machine learning (Chapelle *et al*., 2006; Zhu, 2008). With respect to the CAD system, the semi-supervised learning can learn a hypothesis from a small amount of samples that are carefully diagnosed by medical experts (the labeled data) and then use a large amount of readily available undiagnosed samples (the unlabeled data) to enhance the performance of the learned hypothesis.

Co-training is a prominent semi-supervised classification paradigm proposed by Blum and Mitchell (1998). Two individual classifiers are first trained and then each classifier labels some unlabeled ones for the other to augment the labeled training examples for re-training. The standard co-training assumes two classifiers are respectively trained over two sufficient and redundant views, that is, the attributes should be naturally split into two disjoint subsets, which are conditionally independent given the class and each should be sufficient to train a good classifier. However, this strong assumption can hardly be satisfied in most practical settings. Goldman and Zhou (2000) proposed a revised one named statistical co-training. It relaxes the strong assumptions on two attribute views by using two different type learners both trained by the whole attribute set. However, it requires two supervised learners can partition the instances space into a set of equivalence class and it frequently employs the ten-fold cross validation, quite time-consuming, to estimate the labeling confidence when decides the newly labeled examples and final hypothesis. Therefore, Zhou and Li (2005) proposed the Tri-training algorithm, which uses three classifiers to perform the co-training. It does not require the constraints on attributes, nor does it require the constraints on special classifiers or the cross-validation. Therefore, it has wide applications.

However, many researchers have noted a common problem in co-training-style algorithms that is, the

Corresponding Author: Chao Deng, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

performance of semi-supervised learning are usually not stable because the unlabeled examples may often be wrongly labeled during the co-training process (Blum and Chawla 2001; Vincent and Claire, 2003; Hwa *et al*., 2003; Zhou and Goldman, 2004; Zhou and Li, 2005). Tri-training may suffer more from this problem. How to identify and filter the mislabeled noise data during the co-training process has great significance to enhance the performance of co-training-style approaches.

A promising solution to identify mislabeled noise is exploiting the data editing techniques. For example, Li and Zhou (2005) employed a data editing method to filter the possible mislabeled examples in each iteration of the self-training (a special case of the co-training paradigm) and SETRED is presented. With respect to the Tri-training, the DE-Tri-training, i.e., Tri-training with Data Editing, where the data editing method named Depuration is incorporated into the tri-training process to identify the possible mislabeled ones, then discard or correct them. Both SETRED and DE-Tri-training exploit the data editing by a rote method, that is to say, the data editing would be activated before each re-training without exception. However, the rote method is unwise to gracefully resolve the mislabeling problem under different cases. The experiments by Deng and Guo indicate that when the original labeled data are obviously insufficient and the individual classifier is insufficiently trained, fairly more mislabeled examples are generated and the contribution of data editing to improving the generalization ability is significant; however, when the original labeled data are sufficient and the individual classifier is trained sufficiently, the number of mislabeled data obviously decreases, so the innate error of data editing brings negative effects and directly incurs worse generalization ability than the original algorithm.

Inspired by current booming trend in machine learning, i.e., a cognition model that captures the features of cognitive process behind human learning, may lead to a new learning approach (Mitchell, 2006; Langley, 2007; Zhu *et al*., 2007), this study chooses the participatory learning cognition paradigm proposed by Yager (2004) as a tool to resolve the inevitable mislabeling problem in Tri-training. In detail, the acceptance unit is instantiated as an effective data editing technique named RemoveOnly to filter mislabeled data; and the critic unit is designed as the arousal strategy to detect and inhibit the invalid activation of data editing. Then, each individual classifier of Tri-training is equipped with this specific participatory learning instance and performs co-training process. The revised Tri-training algorithm is called PL-Tri-training (Participatory Learning based Tri-training). It is noteworthy that the activation of data editing, which can not ensure the iterative improvement of classification accuracy, is regarded as invalid. Therefore, the arousal strategy in critic unit contains all the sufficient conditions to ensure the activation of data editing obtaining improvement of accuracy and these sufficient conditions used as criterion to exclude the invalid activation of data editing. Experiments on benchmark medical datasets and the application to medical image detection show the effectiveness of PL-Tri-training for CAD systems.

CO-TRAINING PROCESS OF TRI-TRAINING

The pseudo-code of Tri-training was presented by Zhou and Li (2005). During the co-training process of Tri-training, some unlabeled data are firstly labeled by co-labeling and then these newly labeled data are used to update training set and perform re-training, when some sufficient condition is satisfied. The co-labeling and re-training process for each individual classifier is repeated until none of three individual classifiers changes. The co-labeling and re-training each co-training iteration are showed as:

Co-labeling: Let L and U denote the original labeled and unlabeled set respectively, which are drawn independently from the identical underlying data distribution. In Tri-training, three different classifiers, i.e., H_i, H_j, H_k, are initially trained from three bootstraps of L, respectively. Then, the co-labeling is performed as follows: For every unlabeled data x in U, if H_i and H_j agree on labeling it as $H_i(x)$, then x becomes newly labeled one for the third classifier H_k. Thus, all newly labeled data from U like x are copied into L with new labels and form new candidate training set of H_k. Here, H_i and H_j act as a joint classifier denoted by H_i and H_j. Similarly, the new candidate training sets of H_i and H_j are formed by their corresponding joint classifier.

Because, the new candidate training set might be used to refine the individual classifier in the followed re-training step, if the newly labeled example is wrongly labeled by the joint classifier, the third classifier will obtain a new training example with noise label, which is harmful to its refinement. Therefore, Zhou and Li (2005) derived a sufficient condition to decide whether the new candidate training set should be used for re-training. As a basic condition for detecting the validation of acceptance unit by the critic unit in a participatory learning paradigm, this sufficient condition is formalized to the theorem 1 as following.

Sufficient condition for re-training: The sufficient condition for re-training aims to ensure that the

classification accuracy of individual classifier could be improved by re-training on the new training set.

The sufficient condition is derived from the finding of Angluin and Laird (1988) on the PAC property of hypothesis learned from noisy training examples. That is, the hypothesis minimizing the disagreement with the sequence of training examples will close to the ground-truth hypothesis with the PAC property, if the size m of noisy training set satisfies:

$$m = \frac{c}{\varepsilon^2(1-2\eta)^2} \quad (1)$$

where, c is a constant, ε is the hypothesis worst-case error rate and η (<0.5) is the noise rate on training set. Eq. 1 is reformed as the following utility function:

$$u = \frac{c}{\varepsilon^2} = m(1-2\eta)^2 \quad (2)$$

Obviously, this utility function indicates $u \propto 1/\varepsilon^2$.

According to Eq. (2), the objective of re-training in Tri-training is to ensure the classification error rate ε of hypothesis can be reduced iteratively; meanwhile the size m of new training set for each individual classifier can be iteratively increased.

Let $L_{i,t}$ and $L_{i,t-1}$ denote the newly labeled training subset for H_i from U by the joint classifier H_j and H_k in the t-th and (t-1)-th co-training iteration respectively, where all members of $L_{i,t-1}$ will be put back in U as unlabeled ones in the t-th round. Thus, the training set for H_i in the t-th and (t-1)-th round are $L \cup L_{i,t}$ and $L \cup L_{i,t-1}$, whose sizes are $|L| + |L_{i,t}|$ and $|L| + |L_{i,t-1}|$. Further, let η_L denote the noise rate on the original labeled set L and let $\hat{e}_{i,t}$ (<0.5) denote the error rate upper bound of H_j and H_k on $L_{i,t}$, then the noise rate on $L \cup L_{i,t}$ denoted by $\eta_{i,t}$ could be estimated as:

$$\eta_{i,t} = \frac{\eta_L |L| + \hat{e}_{i,t} |L_{i,t}|}{|L| + |L_{i,t}|} \quad (3)$$

And with Eq. 2, the utility of H_i in t-th iteration denoted by $u_{i,t}$ could be reformed as:

$$u_{i,t} = (|L| + |L_{i,t}|)(1-2\eta_{i,t})^2 \quad (4)$$

Theorem 1: For the successive iterations of Tri-training, i.e., the (t-1)-th and the t-th iteraions where t>1, if

$$0 < \frac{\hat{e}_{i,t}}{\hat{e}_{i,t-1}} < \frac{|L_{i,t-1}|}{|L_{i,t}|} < 1 \quad (5)$$

then the worst error rates $\varepsilon_{i,t-1}$ and $\varepsilon_{i,t}$ of hypotheses, learned by classifier H_i from $L \cup L_{i,t-1}$ and $L \cup L_{i,t}$, satisfy $\varepsilon_{i,t} < \varepsilon_{i,t-1}$.

Theorem 1 can be interpreted as: if the sufficient condition (5) is satisfied, the error rate of classifier H_i would be iteratively reduced meanwhile the size of newly labeled training set is iteratively augmented.

Tri-training employs (5) as the sufficient condition for re-training the classifier H_i in each iteration. Since the detail derivation of Eq. (5) is provided by Zhou and Li (2005), its proof is omitted here.

Although, the sufficient condition (5) can partially help compensate the mislabeling problem, this problem is still serious, especially when the initial labeled set is very limited. In order to more effectively resolve this problem and improve the stability of classification performance, the participatory learning cognition paradigm would be instantiated and equipped into individual classifier H_i.

PARTICIPATORY LEARNING FOR TRI-TRAINING

Participatory learning cognition paradigm is proposed by Yager (2004) and has been used to propose new machine learning approach (Silva *et al*., 2005). Participatory learning paradigm gracefully captures the essential features of human cognition, e.g. the assimilation and accommodation mechanisms in human learning process. The participatory learning paradigm includes two important units prior to the learning updating by new observation data. They are acceptance unit and critic unit. The acceptance unit is exploited to select the proper observation data and pass it to the learning process; however, the critic unit is exploited to supervise the validation of acceptance unit.

Considering that the acceptance unit and critic unit could provide a nature and robust mechanism to filter the inevitable mislabeled data during the co-training process, we employ a data editing named RemoveOnly and design an arousal strategy to instantiate the acceptance unit and critic unit respectively. Thus, when this specific participatory leaning instance is equipped in the three classifiers of Tri-training, the unavoidable mislabeled data could be identified and discarded before re-training; moreover the invalid data editing operation could be detected and inhibited by the arousal strategy in critic unit. Figure 1 shows, the instantiation of participatory learning for individual classifier of Tri-training.

RemoveOnly data editing for acceptance unit: In order to filter the possible mislabeled data in the newly labeled data of each individual classifier, the acceptance unit is instantiated by a specific data editing operation named RemoveOnly, which is the sub-editing operation of Depuration data editing.

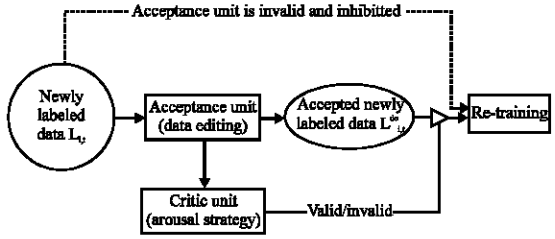

Fig. 1: Participatory learning and its instantiation for individual classifier in Tri-training.

Depuration is the nearest neighbor rule based data editing technique and has been widely used to identify and eliminate the mislabeled and noisy examples in the labeled training set (Sanchez *et al*., 2003). Depuration can be detached into two editing sub-approaches, i.e., RemoveOnly and RelabelOnly. The RemoveOnly approach removes the suspicious examples from the training set, while the RelabelOnly approach only corrects the wrong labels. Jiang and Zhou (2004) showed that the editing effect of RemoveOnly outperforms the Depuration and RelabelOnly. Therefore, the acceptance unit will only employ the RemoveOnly approach to identify and discard the suspicious mislabeled noise in the newly labeled subset $L_{i,t}$. This approach works as follows: for each newly labeled data x in $L_{i,t}$, first, according to the nearest neighbor rule, the k nearest neighbors of x are selected from the training set $L \cup L_{i,t}$; then it checks whether at least k' neighbors in the k nearest of x hold the same class label with x; if not, the concerned example x is identified as a suspicious mislabeled one and to be removed from $L_{i,t}$ together with other suspicious ones in $L_{i,t}$. Sanchez *et al.* (2003) pointed out that when k and k and k' were set to 3 and 2, respectively the Depuration obtained the best editing effect. Since, Jiang and Zhou (2004) took the same settings to reach these findings, the RemoveOnly in acceptance unit adopts this setting too.

Besides, the fact that RemoveOnly can achieve the best editing effect, there are another two merits for employing RemoveOnly. First, when those possibly mislabeled examples are identified, simply removing but not re-labeling them can avoid introducing new noise to the training set. Second, the removing operation makes it simple to quantitatively measure the positive and negative effects of data editing in the followed analysis.

As mentioned before, experiment results have shown that it is unwise to activate the RemoveOnly data editing at each iteration by rote. In participatory learning, the critic unit could act as an inspector of the data editing in acceptance unit and make the activation of data editing appropriate to current case instead of by rote. Therefore, the arousal strategy in critic unit should be designed to measure the performance of data editing in acceptance unit and further detect the validation of data editing so that all the activation of data editing could bring better improvement of accuracy.

Measuring RemoveOnly effects: As shown in theorem 1, the sufficient condition (5) requires the newly labeled training set $L_{i,t}$ satisfy two constraints, i.e., the mislabeling rate $\hat{e}_{i,t}$ of H_i in $L_{i,t}$ will decrease as well as the size $|L_{i,t}|$ will increase. Since, the RemoveOnly data editing in acceptance unit will discard some mislabeled examples from $L_{i,t}$, the effects of RemoveOnly on the quality and the size of $L_{i,t}$ affects whether, the sufficient condition (5) is fulfilled. On the one hand, RemoveOnly exerts positive effect on reducing the mislabeled noise rate $\hat{e}_{i,t}$ by discarding mislabeled examples from $L_{i,t}$; on the other hand, RemoveOnly has a negative effect insofar as it decreases the size $|L_{i,t}|$ by removing suspicious examples. If the editing error of RemoveOnly makes the negative effect more significant than the positive effect, the sufficient condition (5) may become failed, even if it is satisfied before the activation of RemoveOnly. Therefore, it is important to find feasible factors to measure the positive and negative effect of RemoveOnly so that the validation of RemoveOnly under different cases can be decided by an arousal strategy in critic unit.

As shown in Fig. 1, let $L^{de}_{i,t}$ denotes the accepted newly labeled training set after RemoveOnly in acceptance unit discarding some suspicious examples from the newly labeled training set $L_{i,t}$, then the editing effect of RemoveOnly can be measured by two quantitative factors, namely, the size of $L^{de}_{i,t}$ denoted by $|L^{de}_{i,t}|$ and the recall rate of mislabeled data in $L_{i,t}$, which is denoted by $r_{i,t}$ and defined as:

$$r_{i,t} = \frac{\#\text{true mislabeled data removed by acceptance unit}}{\#\text{true mislabeled data in } L_{i,t}}$$

Intuitively, the recall rate $r_{i,t}$ indicates the degree of acceptance unit reducing the mislabeled rate $\hat{e}_{i,t}$ in $L_{i,t}$, meanwhile $|L^{de}_{i,t}|$ reflects the degree to which the $|L_{i,t}|$ is reduced. Thus, the positive and negative effect of RemoveOnly on sufficient condition (5) can accurately be characterized by these two factors, respectively.

Based on these two factors, corresponding to the noise rate $\eta_{i,t}$ on $L \cup L_{i,t}$ computed by Eq. (3), the noise rate on the accepted newly labeled training set $L \cup L^{de}_{i,t}$ denoted by $\eta^{de}_{i,t}$ can be estimated as:

$$\eta^{de}_{i,t} = \frac{\eta_L |L| + (1 - r_{i,t})\hat{e}_{i,t} |L_{i,t}|}{|L| + |L^{de}_{i,t}|} \quad (6)$$

Sufficient conditions for improving generalization: The objective of RemoveOnly in acceptance unit is to improve the generalization ability as much as possible in each re-training iteration. Therefore, it is necessary to further investigate the sufficient conditions that can ensure the improvement of generalization ability under different cases in question, where the accepted newly labeled data become candidate set for re-training.

First, the theorem 1 can be re-interpreted as: if the acceptance unit is invalid and RemoveOnly is not activated in the last iteration, the sufficient condition (5) ensures that the standard re-training process in the current iteration reduces the worst-case error rate as compared with that obtained by last standard re-training.

Corresponding to theorem 1, if the acceptance unit is valid and RemoveOnly is activated in the last iteration, that is, the newly-edited training set $L \cup L^{de}_{i,t-1}$ rather than $L \cup L_{i,t}$ is used in the last re-training iteration, there exists the following theorem 2 for the standard re-training process in current iteration.

Theorem 2: For the successive co-training iterations, i.e., the (t-1)-th and the t-th (t>1) iterations, under the condition that the accepted newly labeled data have been used for re-training in the (t-1)-th iteration, if

$$|L_{i,t}| > |L^{de}_{i,t-1}| > 0 \text{ and } 0 < \frac{\hat{e}_{i,t}}{(1-r_{i,t-1})\hat{e}_{i,t-1}} \le \frac{|L_{i,t-1}|}{|L_{i,t}|} < 1 \quad (7)$$

then the worst-case error rates $\varepsilon^{de}_{i,t-1}$ and $\varepsilon_{i,-1}$ of hypotheses learned by classifier H_i from $L \cup L^{de}_{i,t-1}$ and $L \cup L_{i,t}$, satisfy $\varepsilon_{i,t} < \varepsilon^{de}_{i,t-1}$.

Proof: Given (7), then

$$|L| + |L_{i,t}| > |L| + |L^{de}_{i,t-1}|$$

and

$$0 < \hat{e}_{i,t} |L_{i,t}| \le (1-r_{i,t-1})\hat{e}_{i,t-1} |L_{i,t-1}|$$

Hence,

$$\frac{\eta_L |L| + \hat{e}_{i,t} |L_{i,t}|}{|L| + |L_{i,t}|} < \frac{\eta_L |L| + (1-r_{i,t-1})\hat{e}_{i,t-1} |L_{i,t-1}|}{|L| + |L^{de}_{i,t-1}|}$$

moreover, according to Eq. 3 and 6, the noise rates on $L \cup L^{de}_{i,t-1}$ and $L \cup L^{de}_{i,t}$ are

$$\frac{\eta_L |L| + (1-r_{i,t-1})\hat{e}_{i,t-1} |L_{i,t-1}|}{|L| + |L^{de}_{i,t-1}|}$$

and

$$\frac{\eta_L |L| + \hat{e}_{i,t} |L_{i,t}|}{|L| + |L_{i,t}|}$$

respectively and hence $\eta_{i,t} < \eta^{de}_{i,t-1}$. Thus, combining

$$|L| + |L_{i,t}| > |L| + |L^{de}_{i,t-1}|$$

and

$$\eta_{i,t} < \eta^{de}_{i,t-1}$$

with the two utilities

$$u_{i,t} = (|L| + |L_{i,t}|)(1 - 2\eta_{i,t})^2$$

and

$$u^{de}_{i,t-1} = (|L| + |L^{de}_{i,t-1}|)(1 - 2\eta^{de}_{i,t-1})^2$$

satisfies $u_{i,t} > u^{de}_{i,t-1}$. Since,

$$u \propto \frac{1}{\xi^2}$$

then $\xi_{i,t} < \xi^{de}_{i,t-1}$.

Theorem 2 can be interpreted as follows. When the acceptance unit has been activated in the last re-training iteration, if the condition (7) is satisfied in current iteration, the standard re-training process of Tri-training could make the worst error rate reduced as compared with that obtained in the last re-training iteration by the accepted newly labeled data.

Finally, the theorem 3 shows when the activation of acceptance unit can obtain lower worst-case error rate than that obtained by the standard re-training process in the current same co-training iteration.

Theorem 3: In the t-th (t>1) co-training iteration, $L^{de}_{i,t}$ is obtained from the newly labeled set $L_{i,t}$ by RemoveOnly in the acceptance unit. If,

$$|L^{de}_{i,t}| < |L_{i,t}| \text{ and } r_{i,t} \ge \frac{|L_{i,t}| - |L^{de}_{i,t}|}{2\hat{e}_{i,t} |L_{i,t}|} \quad (8)$$

then the worst-case error rates $\varepsilon_{i,t}$ and $\varepsilon^{de}_{i,t}$ of hypotheses learned by classifier H_i from $L \cup L_{i,t}$ and $L \cup L^{de}_{i,t}$, satisfy $\varepsilon^{de}_{i,t} < \varepsilon_{i,t}$.

Proof: Given $|L^{de}_{i,t}| < |L_{i,t}|$ in (8), then

$$|L| + |L^{de}_{i,t}| < |L| + |L_{i,t}| \quad (9)$$

Given the inequallity

$$r_{i,t} \ge \frac{|L_{i,t}| - |L^{de}_{i,t}|}{2\hat{e}_{i,t} |L_{i,t}|}$$

in Eq. 8, then

$$2r_{i,t}\hat{e}_{i,t} |L_{i,t}| \ge |L_{i,t}| - |L^{de}_{i,t}|$$

Therefore,

$$(2\hat{e}_{i,t} \mid L_{i,t} \mid -2\hat{e}_{i,t} \mid L_{i,t} \mid) + 2r_{i,t}\hat{e}_{i,t} \mid L_{i,t} \mid \geq \mid L_{i,t} \mid - \mid L_{i,t}^{de} \mid$$

and thus,

$$2\hat{e}_{i,t} \mid L_{i,t} \mid -2(1-r_{i,t})\hat{e}_{i,t} \mid L_{i,t} \mid \geq \mid L_{i,t} \mid - \mid L_{i,t}^{de} \mid$$

Interchanging the left term $2\hat{e}_{i,t}|L_{i,t}|$ with right term $|L^{de}_{i,t}|$ yields

$$\mid L_{i,t}^{de} \mid -2(1-r_{i,t})\hat{e}_{i,t} \mid L_{i,t} \mid \geq \mid L_{i,t} \mid -2\hat{e}_{i,t} \mid L_{i,t} \mid$$

and adding |L| to both sides results in:

$$\begin{aligned} &\mid L \mid + \mid L_{i,t}^{de} \mid -2(1-r_{i,t}) \\ &\hat{e}_{i,t} \mid L_{i,t} \mid \geq \mid L \mid + \mid L_{i,t} \mid -2\hat{e}_{i,t} \mid L_{i,t} \mid \end{aligned} \tag{10}$$

On the other hand, according to Eq. 3 and 4,

$$u_{i,t} = \frac{(\mid L \mid + \mid L_{i,t} \mid -2\eta_0 \mid L \mid -2\hat{e}_{i,t} \mid L_{i,t} \mid)^2}{\mid L \mid + \mid L_{i,t} \mid} \tag{11}$$

Meanwhile, according to Eq. 6,

$$u_{i,t}^{de} = \frac{(\mid L \mid + \mid L_{i,t}^{de} \mid -2\eta_0 \mid L \mid -2(1-r_{i,t})\hat{e}_{i,t} \mid L_{i,t} \mid)^2}{\mid L \mid + \mid L_{i,t}^{de} \mid} \tag{12}$$

Thus, following Eq. 10-1 with the assumptions that $\eta_0<0.5$ and $\hat{e}_{i,t}<0.5$, then

$$\begin{aligned} &\mid L \mid + \mid L_{i,t}^{de} \mid -2\eta_0 \mid L \mid -2(1-r_{i,t})\hat{e}_{i,t} \mid L_{i,t} \mid \\ &\geq \mid L \mid + \mid L_{i,t} \mid -2\eta_0 \mid L \mid -2\hat{e}_{i,t} \mid L_{i,t} \mid > 0 \end{aligned} \tag{13}$$

Finally, according to Eq. 9, 13, 11 and 12, then $u^{de}_{i,t}>u_{i,t}$. Since, $u \propto 1/\varepsilon_2$, $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}$ is proved.

Theorem 3 can be interpreted as follows. Regardless of whether, the acceptance unit (i.e., RemoveOnly) is activated or not in the last iteration and whether the newly labeled data before RemoveOnly editing could reduce the error rate or not in the current iteration, when the condition (8) is satisfied, the activation of acceptance unit obtains a better hypothesis with higher classification accuracy in the current iteration than the standard co-training process of Tri-training. Note that this does not indicate that the activation of acceptance unit would certainly reduce the worst-case error rate as compared with that in the last iteration.

In summary, theorem 1 and 2 show the conditions sufficient for the iterative reduction of error rate in the current iteration when the acceptance unit is inhibited or activated in the last iteration; theorem 3 shows the sufficient condition for activating acceptance unit in order to obtain a more accurate hypothesis in the current iteration.

Arousal strategy for critic unit: Theorems 1-3 investigate all the possible cases that might be encountered when acceptance unit filters the newly labeled data using RemoveOnly in the t-th iteration. The arousal strategy in critic unit employs these sufficient conditions in the theorems as elemental criterion to detect that the acceptance unit is valid or invalid. As shown in Fig. 1, if the acceptance unit is detected as invalid, its activation would be inhibited and Tri-training performs standard re-training. In detail, the arousal strategy in critic unit is designed as follows:

In the t-th iteration,

- If acceptance unit is not activated in the (t-1)-th iteration and $|L_{i,t}|>|L_{i,t-1}|$ holds, in the case that both the sufficient condition for $\varepsilon_{i,t}<\varepsilon_{i,t-1}$ in theorem 1 and the sufficient condition for $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}$ in theorem 3 are satisfied, i.e., $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}<\varepsilon_{i,t-1}$, then the current acceptance unit is valid and will be activated
- If acceptance unit has been activated in the (t-1)-th iteration and $|L_{i,t}|>|L_{i,t-1}|$ holds, in the case that both the sufficient condition for $\varepsilon_{i,t}<\varepsilon_{i,t-1}$ in theorem 2 and the sufficient condition for $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}$ in theorem 3 are satisfied, i.e., $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}<\varepsilon_{i,t-1}$, the current acceptance unit is valid and will be activated
- Except for the above cases, the acceptance unit is regarded as invalid and will be inhibited.

Strategies (1) and (2) show that after the standard Tri-training improves the classification accuracy of hypothesis in the current iteration, the critic unit will activate the RemoveOnly in acceptance unit in order to achieve as much improvement as possible. In addition, according to strategies (1) and (3), if theorem 1 are satisfied at the same time, i.e., $\varepsilon_{i,t}<\varepsilon_{i,t-1}$ is met, then RemoveOnly will not be activated unless theorem 3 is further satisfied, i.e., $\varepsilon^{de}_{i,t}<\varepsilon_{i,t}$. Similarly, according to strategies (2) and (3), if theorem 2 are satisfied, i.e., $\varepsilon_{i,t}<\varepsilon^{de}_{i,t-1}$ is met, then RemoveOnly will not be activated unless theorem 3 is further satisfied. This implies that upon updating the training set for re-training, the set $L^{de}_{i,t}$ generated by RemoveOnly editing in acceptance unit has lower priority than the set $L_{i,t}$, which was originally labeled by the basic co-training process in Tri-training. Accordingly, $L^{de}_{i,t}$ will only be employed for the cases shown by strategies (1) to (2), for which it can provide the greater improvement in accuracy than can $L_{i,t}$. This prevents the possibility that the activation of acceptance unit results in the worse performance than the standard co-training in Tri-training.

THE PL-TRI-TRAINING ALGORITHM

The PL-Tri-training equips each individual classifier by the participatory learning and the standard co-training process is extended. First, the newly labeled training set for each individual classifier is generated through the co-labeling step as standard Tri-training does. Then, the acceptance unit filters the mislabeled ones from the newly labeled data by the RemoveOnly data editing and the accepted newly labeled data are obtained. Finally, the critic unit employs the arousal strategy to decide whether, the acceptance unit is valid or invalid for improving the accuracy according to current cases. If valid, then the acceptance unit is activated and the accepted newly labeled data are used for re-training; if invalid, the acceptance unit is inhibited and the newly labeled data will bypass the acceptance unit and are directly checked by the sufficient condition (5) the same as the standard.

Table 1 presents, the pseudo-code of PL-Tri-training algorithm. The functions Measure error (H_j, H_k, L) and Measure Recall (H_j and H_k, L) estimate the error rate of the joint classifier H_j and H_k and the recall rate of RemoveOnly on the newly labeled set generated by H_j and H_k, respectively. Since, it is difficult to estimate the ground-truth error rate and recall rate on the unlabeled set, based on the assumption that unlabeled set U and original labeled set L are drawn from the identical underlying distribution, here the original labeled set L is employed for estimating the measures. In detail, the error rate of H_j and H_k on $L_{i, t}$, which is labeled by H_j and H_k from U, is approximated by the error rate of H_j and H_k on the subset that is labeled by H_j and H_k from L with same labels. Similarly, the recall rate of RemoveOnly on L_i is estimated by the recall rate on the same subset from L instead. Note that the function RemoveOnly (L_i) performs data editing by identifying and discarding the mislabeled examples in L_i as described before. The final hypothesis H(x) is also based on the majority voting of three final individual classifiers.

Table 1: Pseudo-code for PL-Tri-training algorithm

Input: L: Original labeled set; U: Unlabeled set;
Learn: supervised learning algorithm
for $i \in \{1..3\}$ do //initialize three individual classifiers
$S \leftarrow$ Bootstrapsample (L)
$H_i \leftarrow$ Leans (S_i)
$e_i' \leftarrow 0.5; l_i' \leftarrow 0; lde_i' \leftarrow 0; r_i' \leftarrow 0; valid_i' \leftarrow$ FALSE
repeat until none of H_i ($i = \{1..3\}$) changes // co-training process
for $i \in \{1..3\}$ do
$L_i \leftarrow \phi; LDE_i \leftarrow \phi; update_i \leftarrow$ FALSE; $valid_i \leftarrow$ FALSE
$e_i \leftarrow$ MeasureError(H_j and H_k, L)
$r_i \leftarrow$ MeasureRecall(H_j and H_k, L)
for every $x \in U$ do // co-labeling step
if $H_j(x) = H_k(x)$ $(j,k \neq i)$ then $L_i \leftarrow L_i \cup \{(x, H_j(x))\}$
end for
$LDE_i \leftarrow$ RemoveOnly // data editing in acceptance unit
if $valid_i' =$ FALSE //acceptance unit is invalid last iteration
if condition (5) in theorem 1 is satisfied // strategy (1)
$update_i \leftarrow$ TRUE
if condition (8) in theorem 3 is satisfied
$valid_i \leftarrow$ TRUE
if $valid_i' =$ TRUE//acceptance unit is activated last iteration
if condition (7) in theorem 2 is satisfied // strategy (2)
$update_i \leftarrow$ TRUE
if condition (8) in theorem 3 is satisfied
$valid_i \leftarrow$ TRUE
for $i \in \{1..3\}$ do // re-training step
if $valid_i =$ TRUE //re-train by accepted newly labeled data
$H_i \leftarrow$ Learn($L \cup LDE_i$)
$e_i' \leftarrow e_i; l_i' \leftarrow |L|; valid_i' \leftarrow valid_i; r_i' \leftarrow r_i; lde_i' \leftarrow |LDE_i|$
else if $update_i =$ TRUE// re-train as standard tri-training
$H_i \leftarrow$ Learn($L \cup L_i$)
$e_i' \leftarrow e_i; l_i' \leftarrow |L|; valid_i' \leftarrow$ FALSE
end of repeat
Output: $H(x) \leftarrow \arg\max_{y \in label} \sum_{i:H_i(x)=y} 1$

EXPERIMETNS ON BENCHMARK DATASETS

In order to verify the effectiveness of the PL-Tri-training on medical diagnosis tasks, three benchmark medical diagnosis datasets from UCI repository (Blake *et al*., 1998) are used. Information on these datasets is tabulated in Table 2. The pos/neg presents the percentage of positive data against that of negative ones.

Experiment settings: For each data set, 25% data are kept aside as test set to evaluate the performance of the learned hypothesis, while the remaining 75% data are training set. The training set is partitioned into original labeled set L and unlabeled set U according to different unlabelled rates, i.e., 80 and 60%. For instance, among 1,000 training examples, when the unlabelled rate is 80%, 200 examples are put into L with labels and 800 examples are put into U without labels. In principle, the pos/neg ratio on L, U and test set are identical to that on the original data set.

Under each unlabelled rate, three different random partitions of L and U are generated and one independent run is performed for each random partition, then the average classification error rate of three runs is computed and as the final performance evaluation.

The performance of PL-Tri-training is compared with other two tri-training algorithms, i.e., the standard

Table 2: Medical diagnosis datasets from UCI

Dataset	Attribute	Size	Class	Pos/neg
Diabetes	8	768	2	65.1/34.9%
Hypothyroid	25	3,163	2	4.8/95.2%
Wdbc	30	569	2	37.3/62.7%

Table 3: Average error rates under unlabeled rate 90%

		PL-Tri-training		AC-Tri-training		Tri-training	
Datasets	Initial	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)
Diabetes	0.305	0.243	20.7	0.279	8.53	0.284	6.9
Hypothyroid	0.093	0.052	44.1	0.052	44.10	0.064	31.2
Wdbc	0.108	0.073	32.3	0.094	13.00	0.098	9.3
Average	0.169	0.123	32.4	0.142	21.70	0.146	15.8

Table 4: Average error rates under unlabeled rate 70%

		PL-Tri-training		AC-Tri-training		Tri-training	
Datasets	Initial	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)
Diabetes	0.281	0.241	14.2	0.260	7.5	0.262	6.8
Hypothyroid	0.057	0.033	42.1	0.061	-7.0	0.041	28.1
Wdbc	0.073	0.057	22.0	0.069	6.2	0.069	4.9
Average	0.137	0.110	26.1	0.124	2.2	0.124	13.3

Tri-training and AC-Tri-training. The standard Tri-training is the same as the original one (Zhou and Li, 2005). AC-Tri-training only equips the individual classifiers with the acceptance unit but without the arousal strategy of critic unit and the RemoveOnly operation in acceptance unit is activated at each re-training iteration by rote.

The BP neural network in WEKA toolkits (Witten and Frank, 2005) is used as the learning algorithm for each individual classifier. The Heterogeneous Value Difference Metric (HVDM) (Wilson and Martinez, 1997) is used as the similarity metric between two examples for the RemoveOnly in acceptance unit.

Experiment results: Table 3 and 4 summarize, the average error rates of three tri-training algorithms under different unlabeled rate for each data set. The column initial presents the error rate of the hypothesis at round 0, i.e., the combination of the three initial individual classifiers trained from L; the column final presents the error rate of the final hypothesis generated by three final individual classifiers when co-training process finished; the column improve presents the improvement of the final hypothesis over the initial hypothesis and is computed by the reduction percentage between columns final and initial; the row avg. in each table shows the average results over all the data sets. For each data set row, the biggest improvement percentage has been boldfaced. Note that some values in the tables may look inconsistent due to truncation. For example, the final column of AC-Tri-training and Tri-training on wdbc appear identical in Table 4 but the improvement is not equal.

Table 3 and 4 show that PL-Tri-training algorithm can effectively improve the performance of hypothesis under all data sets and all unlabeled rates. If the improvements of each algorithm are averaged across over all the data sets and unlabeled rates, it can be found that the average improvement of PL-Tri-training is the most significant, i.e., 29.3%, while that of Tri-training and AC-Tri-training are

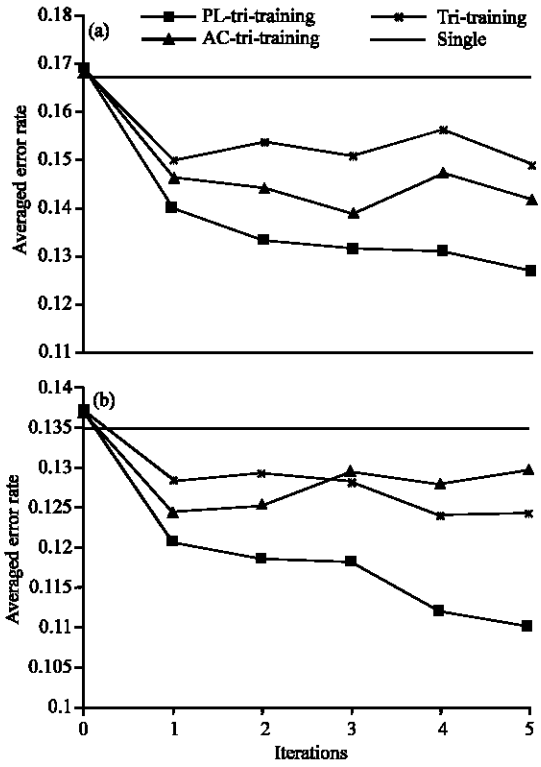

Fig. 2: Iterative change of average error rate. a) Unlabeled rate 90%; b) Unlabeled rate 70%

14.5 and 11.9%. This proves that PL-Tri-training can effectively improve the performance of Tri-training; however the AC-Tri-training reduces the performance of Tri-training. This obvious difference shows that the arousal strategy of critic unit in PL-Tri-training has great significance to improve the performance of Tri-training.

Moreover, Table 3 and 4 also show that if the algorithms are compared through counting the number of winning data sets, i.e., the number of data sets on which one algorithm achieves the best improvement among compared algorithms, PL-Tri-training is always the winner. In detail, under 90% unlabeled rate, PL-Tri-training has 3 winning sets, while AC-Tri-training and Tri-training only have 1 winning sets and zero wining set respectively; under 70% unlabeled rate, PL-Tri-training has 3 winning sets ,while AC-Tri-training and Tri-training only have zero winning set.

Figure 2 depicts, the error rates change of the compared algorithms along with the co-training iterations. Besides three tri-training semi-supervised learning algorithms, on each data set the supervised individual classifier based on BPNN is trained only from the original labeled data set L and its average error rates are denoted by horizon lines named single. Note that the error rates

have been averaged across all the data sets and since the semi-supervised algorithms may terminate in different iterations, the error rates at termination are used as the error rates of the iterations after termination.

The Fig. 2a and b reveal that the final hypotheses generated by semi-supervised algorithms have better performance than that generated by the individual classifier trained by BPNN supervised learning. All subfigures also confirm that three tri-training semi-supervised learning algorithms can exploit the unlabeled data to reduce the error rate from the initial hypothesis. Moreover, it is obvious that at each iteration the performance of PL-Tri-training significantly outperforms the standard Tri-training and AC-Tri-training. However, the performance of AC-Tri-training is worse than the standard Tri-training at some iterations, which proves the significance of arousal strategy in critic unit again. In addition, the performance curve of each algorithm reveals that the average error rate of PL-Tri-training is almost continuous reduced till the termination, by contrast, the fluctuations of Tri-training and AC-Tri-training are observed in Fig. 2a and b. This implies that the inevitable mislabeling problem has minimal influence to PL-Tri-training.

APPLICATION TO SMALL PULMONARY NODULES DETECTION IN CHEST CT IMAGES

The detection of small pulmonary nodules is a huge clinical challenge for radiologists in diagnosing various lung diseases. Small pulmonary nodules are a common part of the clinical profile of many lung diseases, including the lung cancer, which is the leading cause of cancer death among both men and women. In high-resolution CT chest images, nodules larger than 10 mm are easy to detect, but smaller pulmonary nodules less than 10mm frequently occur. The precise detection of these smaller nodules is laborious and tedious for radiologists. Especially, in the central lung regions, nodules are confused with blood vessels and so contiguous CT slices have to be evaluated to discriminate between spherical nodules and tubular vessels structures. This is a time-consuming and error-susceptible task. Since, many CT slices per patient must be interpreted for precise detection, the detection of small pulmonary nodules is extremely intensive. Therefore, the reliable automated detection of small pulmonary nodules on CT scans is important for the early detection of lung cancer and other nodular lung diseases (Das *et al*., 2006).

Application settings: The clinical data set used here includes 125 high-resolution CT chest images collected by the 2nd Affiliated Hospital of Harbin Medical University and each image has 1024×1024 resolution and 12 bits per pixel. Among the total 125 images, 15 images contain at least one pulmonary nodule marked by radiologists, while the other 110 images are unmarked. Each image is split into a set of 32×32 blocks. For each 32×32 block, the five features (Jia *et al*., 2006; Li and Zhou, 2007), namely, the average density, density variance, energy variance, block activity and spectral entropy, are extracted. Furthermore, for each block fragmented from a marked image, if the block contains at least 100 pixels of a marked pulmonary nodule, it is labeled as a positive example; otherwise, it is labeled as a negative one. All the blocks fragmented from unmarked images are regarded as unlabeled data. After removing blocks ineligible to be marked (for example, they may depict bodily regions outside the lungs), the data set consisted of 54 positive examples, 112 negative examples and an additional 1493 unlabeled examples. Thus, the task is to detect, whether a block contains small pulmonary nodules. It is noteworthy that the unlabeled rate in this task is close to 90%.

Further, Fig. 3 shows how the PL-Tri-training obtains new positive blocks with pulmonary nodules in an original unmarked image. Figure 3a is the original unmarked CT image with 512×512 resolutions and Fig. 3b is the corresponding lung parenchyma image after removing background. The right side in Fig. 3b shows, when the acceptance unit is allowed to be activated in certain co-training iteration, three 32×32 blocks containing small pulmonary nodules are newly labeled as positives and added into training set, by contrast, some other newly labeled but suspicious positive blocks are discarded by RemoveOnly data editing in the acceptance unit.

In order to provide fair comparison between PL-Tri-training and other algorithms, five-fold cross validation is performed (Witten and Frank, 2005). In detail, the labeled data are firstly partitioned into 5 folds and the class distribution in each fold is similar to that in the original labeled data. Then, each fold is in turn selected as the test set and the other four folds serve as the labeled training set combining with the unlabeled data to train classifiers. Finally, the performance values on five test folds are averaged as the result of five-fold cross validation. We repeat the five-fold cross validation 3 times and obtain average performance.

Performance comparison: Two clinical factors, false negative rate and false positive rate, can be used to evaluate the detection accuracy. According to the clinical experience, misclassifying the blocks with pulmonary nodules as normal ones will delay the accurate diagnosis and lessen the optimal treatment chances. Therefore, the

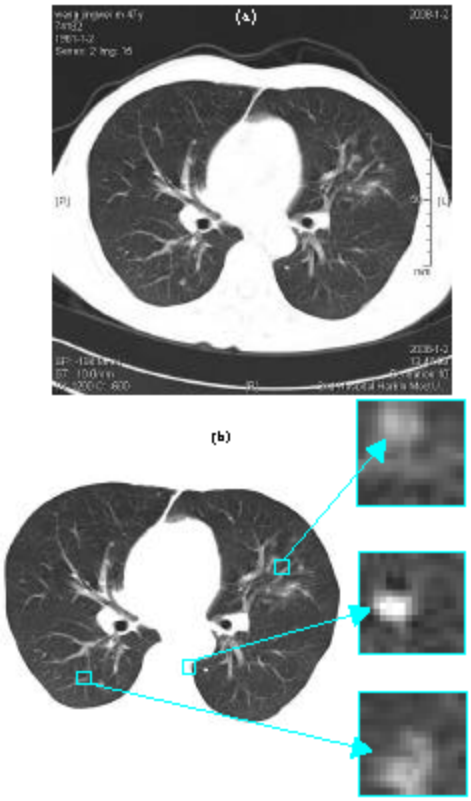

Fig. 3: Using unlabeled examples and acceptance unit to label and filter blocks in unmarked CT image. (a) the original unmarked chest CT image; (b) the lung parenchyma and the accepted newly labeled positive blocks with pulmonary nodules after RemoveOnly data editing in acceptance unit is activated

false negative rate is used as a major factor for evaluating any automated pulmonary nodules detection algorithm, which is defined as the number of positive examples (i.e., blocks with nodules) misclassified as negative ones divided by the total number of ground-truth positive examples. Meanwhile, since the doctors make diagnosis according to the blocks detected by the algorithm, misclassifying the normal blocks as blocks with pulmonary nodules will increase the burden on the doctors. Therefore, the false positive rate is used as another evaluation factor and is defined as the number of negative examples (i.e., normal blocks without nodule) misclassified as positive divided by the total number of examples classified as positive.

The average false negative rate and the average false positive rate of all the algorithms against iteration are plotted in Fig. 4 and 5; the BPNN trained only on the labeled data serves as the baseline for comparison.

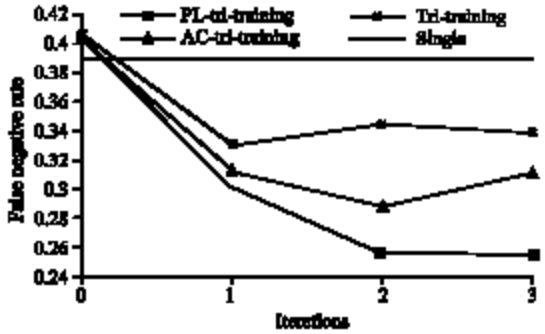

Fig. 4: Iterative change of average false negative rates

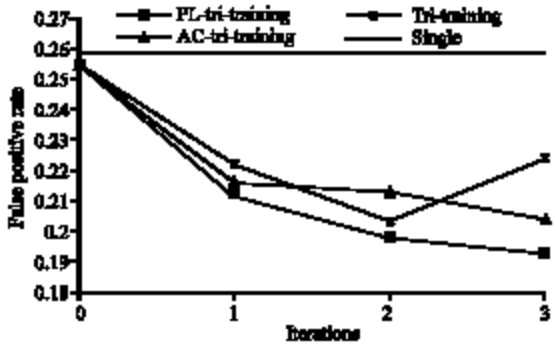

Fig. 5: Iterative change of average false positive rates

Figure 4 and 5 show that as compared with the BPNN baselines, the three semi-supervised learning algorithms generate better hypotheses by exploiting unlabeled examples. Furthermore, these figures reveal that the PL-Tri-training algorithm benefits most from the unlabeled examples and its final hypothesis clearly outperforms those learned by AC-Tri-training and Tri-training. After three learning iterations, the initial average negative false rate in Fig. 4 is decreased from the initial value of 0.402-0.254, 0.311 and 0.338 by PL-Tri-training, AC-Tri-training and Tri-training, respectively. It is quite impressive that PL-Tri-training reduces the average false negative rate by 36.9%, as compared with 23.0% for Tri-training and 16.2% for AC-Tri-training. Meanwhile, after three iterations, the initial average negative positive rate in Fig. 5 is decreased from the initial value of 0.254-0.197, 0.204 and 0.224 by PL-Tri-training, AC-Tri-training and Tri-training, respectively. The largest reduction of 24.2% is also generated by PL-Tri-training, as compared with 12.0% generated by Tri-training and 19.6% generated by AC-Tri-training.

For clinical diagnoses, PL-Tri-training's relatively large reduction of both the false negative rate and the false positive rate suggests that this algorithm classifies much fewer normal blocks as positive ones. More importantly, it is able to more effectively use the unlabeled

examples to detect more ground-truth pulmonary nodules than Tri-training and AC-Tri-training. In other words, at an unlabeled rate of 90% even if the radiologists only focus on a limited marking of 10% of small pulmonary nodules in chest CT images, the CAD system based on PL-Tri-training can effectively detect more ground-truth small pulmonary nodules without misclassifying more normal blocks than Tri-training and AC-Tri-training.

CONCLUSION

In this study, a participatory learning cognition paradigm based Tri-training algorithm named PL-Tri-training algorithm is proposed. The acceptance unit instantiated as the RemoveOnly data editing can effectively identify and discard the mislabeled examples from the newly labeled data and the critic unit designed as arousal strategy can avoid the invalid data editing in acceptance unit.

Experiments on three medical diagnosis benchmark datasets show that the two units of participatory learning both are indispensable to effectively resolve the mislabeling problem in the co-training process of Tri-training, hence stably improve the generalization ability of final hypothesis under different practical cases. Application to the detection of small pulmonary nodules in CT images proves that PL-Tri-training can effectively reduce the false negative rate and false positive rate for clinical diagnosis.

Moreover, since the participatory learning paradigm provides an effective mechanism to filter the mislabeled examples and avoid the risk of filtering by rote, it could be exploited to resolve the inevitable mislabeling problem in other co-training-style semi-supervised algorithms used in computer aided diagnosis systems.

REFERENCES

Anagnostopoulos, I. and I. Maglogiannis, 2006. Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances. Med. Biol. Eng. Comput., 44: 773-784.

Angluin, D. and P. Laird, 1988. Learning from noisy examples. Machine Learning, 2 (4): 343-370.

Blake, C., E. Keogh and C.J. Merz, 1998. UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, http://www.ics.uci.edu/~mlearn/ML Repository.html.

Blum, A. and T. Mitchell, 1998. Combining labeled and unlabeled data with co-training. In: Proc. 11th Annu. Conference on Computational Learning Theory. Wisconsin, USA, pp: 92-100.

Blum, A. and S. Chawla, 2001. Learning from labeled and unlabeled data using graph mincuts. In Proc. 18th International Conference on Machine Learning (ICML). Williamstown, MA., pp: 19-26.

Chapelle, O., B. Schoelkopf and A. Zien, 2006. Semi-supervised Learning. MA: MIT Press, Cambridge.

Das, M., G. Muhlenbruch, A.H. Mahnken, T.G. Flohr, L. Gundel, S. Stanzel, T. Kraus, R.W. Gunther and J.E. Wildberger, 2006. Small pulmonary nodules: Effect of two computer-aided detection systems on radiologist performance. Radiology, 241 (2): 564-571.

Goldman, S. and Y. Zhou, 2000. Enhancing supervised learning with unlabeled data. In: Proc. 17th International Conference on Machine Learning (ICML), San Francisco, CA, pp: 327-334.

Hwa, R., M. Osborne, A. Sarkar and M. Steedman, 2003. Corrected cotraining for statistical parsers. In: Proc. 20th International Conference on Machine Learning (ICML) Workshop on Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. Washington, DC, pp: 95-102.

Jia, X.H., Z. Wang and S.C. Chen, 2006. Fast screening out true negative regions for microcalcification detection in digital mammograms. Trans. Nanjing Univ. Aeronautics and Astronautics, 23 (1): 52-58.

Jiang, Y. and Z.H. Zhou, 2004. Editing training data for kNN classifiers with neural network ensemble. In: Proc. IEEE. Int. Sym. Neural Networks (ISNN). Dalian, China, pp: 356-361.

Li, M. and Z.H. Zhou, 2005. SETRED: Self-training with editing. In: proc. 9th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD). Hanoi, Vietnam, pp: 611-621.

Li, M. and Z.H. Zhou, 2007. Improve Computer Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples. IEEE. Trans. Syst. Man and Cybernetics-PART A, 37 (6): 1088-1098.

Langley, P., 2007. Artificial Intelligence and Cognitive Systems. In: Cohen, P. (Ed.). AI: The First Hundred Years. Menlo Park, CA: AAAI Press.

Mitchell, T., 2006. The discipline of machine learning. Carnegie Mellon University, Technology Report CMU-ML-06-108.

Sanchez, J.S., R. Barandela, A.I. Marqués, R. Alejo and J. Badenas, 2003. Analysis of new techniques to obtain quality training sets. Pattern Recognition Lett., 24 (7): 1015-1022.

Silva, L., F. Gomide and F.R. Yager, 2005. Participatory Learning in Fuzzy Clustering. In: Proc. The 14th IEEE International Conference on Fuzzy Systems, pp: 857-861.

Vincent, N. and C. Claire, 2003. Bootstrapping coreference classifiers with multiple machine learning algorithms. In: Proceeding Conference Empirical Methods in Natural Language Processing. Sapporo, Japan, pp: 113-120.

Wilson, D.R. and T.R. Martinez, 1997. Improved heterogeneous distance functions. J. Artificial Intell. Res., 6 (1): 1-34.

Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 2nd Edn. San Francisco, CA: Morgan Kaufmann.

Yager, R.R., 2004. Participatory learning: A paradigm for building better digital and human agents. Law, Probability and Risk 3, Oxford, pp: 133-145.

Zhou, Y. and S. Goldman, 2004. Democratic co-learning. In: Proc.16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, FL, pp: 594-602.

Zhou, Z.H. and M. Li, 2005. Tri-training: exploiting unlabeled data using three classifiers. IEEE. Trans. Knowledge and Data Eng., 17 (11): 1529-1541.

Zhu, X.J., 2008. Semi-supervised learning literature survey. University of Wisconsin-Madison, Wisconsin, Technology Report and Computer Sciences, TR1530.

Zhu, X.J. *et al*., 2007. Humans Perform Semi-Supervised Classification Too. In: 22nd AAAI Conference on Artificial Intelligence (AAAI). CA: AAAI Press.