Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples

Ming Li and Zhi-Hua Zhou, *Senior Member, IEEE*

***Abstract*— In computer-aided diagnosis, machine learning techniques have been widely applied to learn hypothesis from diagnosed samples in order to assist the medical experts in making diagnosis. To learn a well-performed hypothesis, a large amount of diagnosed samples are required. Although the samples can be easily collected from routine medical examinations, it is usually impossible for the medical experts to make diagnosis for each of the collected samples. If hypothesis could be learned in presence of a large amount of undiagnosed samples, the heavy burden on the medical experts could be released. In this paper, a new semi-supervised learning algorithm named *Co-Forest* is proposed. It extends the *co-training* paradigm by using a well-known ensemble method named *Random Forest*, which enables Co-Forest to estimate the labeling confidence of undiagnosed samples and produce the final hypothesis easily. Experiments on benchmark data sets verify the effectiveness of the proposed algorithm. Case studies on three medical data sets and a successful application to microcalcification detection for breast cancer diagnosis show that undiagnosed samples are helpful in building computer-aided diagnosis systems, and Co-Forest is able to enhance the performance of the hypothesis learned on only a small amount of diagnosed samples by utilizing the available undiagnosed samples.**

***Index Terms*— Computer-aided diagnosis, machine learning, semi-supervised learning, co-training, ensemble learning, random forest, microcalcification cluster detection**

I. Introduction

Machine learning techniques have been successfully applied to computer-aided diagnosis (CAD) systems [20] [35] [42]. These methods learn hypotheses from a large amount of diagnosed samples, i.e. the data collected from a number of necessary medical examinations along with the corresponding diagnoses made by medical experts, in order to assist the medical experts in making diagnosis in future.

To make the CAD systems perform well, a large amount of samples with diagnosis are required for learning. Usually these samples can be easily collected from routine medical examinations. However, making diagnosis for such a large amount of cases one by one places heavy burden on medical experts. For instance, to construct a CAD system for breast cancer diagnosis, radiologists have to label every *focus* in a huge amount of easily obtained high resolution mammograms. This process is usually quite time-consuming and inefficient. One possible solution is to learn hypothesis from a small amount of samples that are carefully diagnosed by medical experts (the labeled data) and then utilize a large amount of readily available undiagnosed samples (the unlabeled data) to enhance the performance of the learned hypothesis. In machine learning, this technique is called *learning with labeled and unlabeled data*.

An effective way to enhance the performance of the learned hypothesis by using the labeled and unlabeled data together is known as *semi-supervised learning* [8] [32] [46], where an initial hypothesis is usually learned from labeled data and then refined with the information derived from the unlabeled ones. *Co-training* [4] is an attractive semi-supervised learning paradigm, which trains two classifiers through letting them label the unlabeled examples for each other. In co-training the data should be described by two *sufficient* and *redundant* attribute subsets, each of which is sufficient for learning and independent to the other given class label.

Although co-training has already been successfully applied to some fields [4] [25] [30], the requirement on two sufficient and redundant attribute subsets is too strong to be met in many real-world applications. Goldman and Zhou [17] extended co-training by replacing the requirement on two sufficient and redundant attribute subsets with the requirement on two different supervised learning algorithms whose hypotheses partition the instance space into a set of equivalence classes. Ten-fold cross validation is frequently applied to find the confident examples to label in every training iteration and produce the final hypothesis, which makes both the learning process and prediction time-consuming.

In this paper, a new co-training style algorithm named Co-Forest, i.e. CO-trained random FOREST, is proposed. It extends the co-training paradigm by incorporating a well-known ensemble learning [13] algorithm named *Random Forest* [7] to tackle the problems of how to determine the most confident examples to label and how to produce the final hypothesis. Since ensemble learning has been successfully applied to many medical problems [35] [41] [42], the particular settings enables Co-Forest to exploit the power of ensemble for better performance of the learned hypothesis in semi-supervised learning. Since Co-Forest requires neither the data be described by sufficient and redundant attribute subsets nor special learning algorithms which frequently employ time-consuming cross validation in learning, it could be easily applied in CAD systems. Experiments on UCI data sets verify the effectiveness of the proposed algorithm. Case studies on three medical diagnosis tasks and a successful application to microcalcifi-

Manuscript received March 5, 2006; revised October 23, 2006; accepted February 12, 2007. This work was supported by the National Science Foundation of China (60325207, 60635030), the Jiangsu Science Foundation Key Project (BK2004001), the Foundation for the Author of National Excellent Doctoral Dissertation of China (200343), and the Graduate Innovation Program of Jiangsu Province.

The authors are with National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China. (E-Mail: {lim, zhouzh}@lamda.nju.edu.cn)

cation cluster detection in digital mammograms show that the undiagnosed samples are beneficial and the hypothesis learned by Co-Forest achieves remarkable performance, even though it is learned from a large amount of undiagnosed samples in addition to only a small amount of diagnosed ones. Hence, constructing CAD system with Co-Forest may release the burden on medical experts for diagnosing a large number of samples.

The rest of the paper is organized as follows: Section II briefly reviews semi-supervised learning and ensemble learning. Section III presents Co-Forest. Section IV reports the experimental results on UCI data sets and case studies on three medical diagnosis data sets. Section V describes the application to microcalcification cluster detection in digital mammograms. Finally, Section VI concludes the paper.

II. Background

A. Semi-Supervised Learning

In traditional supervised learning, all training data should be labeled before learning, and classifiers are then trained on these labeled data. When a portion of the training data are unlabeled, an effective way to combining labeled and unlabeled data in learning is known as *semi-supervised learning* [8] [32] [46], where an initial hypothesis is firstly learned from the labeled data and then refined through the unlabeled ones labeled by certain automatic labeling strategy.

Many semi-supervised learning algorithms have been proposed. Typical ones include using EM [12] approach to estimate the parameters of a generative model and the probability of unlabeled examples being in each class [26] [28] [34]; constructing a graph on training data by certain similarity between examples and imposing label smoothness on the graph as a regularization term [3] [38] [47]; using a transductive inference for support vector machines on a special test set [23]; etc.

A preeminent work in semi-supervised learning methods is the *co-training* paradigm proposed by Blum and Mitchell [4]. In co-training, two classifiers are trained on two sufficient and redundant sets of attributes respectively. Each classifier labels several unlabeled examples whose labels are most confidently predicted from its point of view. These newly labeled examples are used to augment the labeled training set of the other classifier. Then, each classifier is refined with its augmented labeled training set. They [4] showed that any weak hypothesis could be boosted from the unlabeled data if the data meet the class-conditional independent requirement and the target concept is learnable with random classification noise. Dasgupta et al. [11] derived a generalization error bound for the co-trained classifier, which indicates that when the requirement on the existence of sufficient and redundant attribute subsets is met, the co-trained classifiers can make fewer generalization errors by maximizing their agreements over the unlabeled data.

However, although co-training has been applied in some applications such as visual detection [25], noun phrase identification [29] and statistical parsing [21] [30] [36], the requirement on sufficient and redundant attribute subsets can be hardly met in most real-world applications. Goldman and Zhou [17] relaxed this constraint on data by using two supervised learning algorithms, each of which produces hypothesis that is able to partition the instance space into a set of equivalence classes. Recently, through using three classifiers instead of two classifiers, Zhou and Li [43] proposed the tri-training algorithm, which requires neither sufficient and redundant attribute subsets nor special supervised learning algorithms that could partition the instance space into a set of equivalence classes. Another variant of co-training involving multiple classifiers has been presented by Zhou and Goldman [39]. It is worth mentioning that co-training paradigm is not only applicable to classification tasks. Recently, a co-training style algorithm for semi-supervised regression has been proposed [44], which does not require sufficient and redundant attribute subsets.

B. Ensemble Learning

Ensemble learning paradigms train multiple component learners and then combine their predictions. Ensemble techniques can significantly improve the generalization ability of single learners, and therefore ensemble learning has been a hot topic during the past years [13].

An ensemble is usually built in two steps. The first step is to generate multiple component classifiers and the second step is to combine their predictions. According to the way to generate component classifiers, current ensemble learning algorithms fall into two categories, i.e., algorithms that generate component classifiers in parallel and algorithms that generate component classifiers in sequence. Bagging [5] is a representative of the first category. It generates each classifier on an example set bootstrap sampled [14] from the original training set in parallel, and then combines their predictions using majority voting. Other well-known algorithms in this category include Random Subspace [19], Random Forest [7], etc. In the second category the representative algorithm is AdaBoost [15], which sequentially generates a series of classifiers on the data set by making the subsequent classifier focus on the training examples misclassified by the former classifiers. Other well-known algorithms in this category include Arc-$x4$ [6], LogitBoost [16], etc.

Ensemble learning has already been successfully applied to computer-aided diagnosis. Representative applications include employing a two-level ensemble to identify lung cancer cells in the images of the specimens of needle biopsies obtained from the bodies of the subjects to be diagnosed [42]; employing an ensemble to reduce the high prediction variance exhibited by a single classifier in predicting the outcome in In-Vitro Fertilisation [10]; employing an ensemble for breast cancer diagnosis, where the ensemble is adapted to the required sensitivity and specificity by manipulating the proportion of the benign samples to the malignant samples in training data [35]; employing an ensemble for the classification of glaucoma by using the Heidelberg Retina Tomograph to derive the measurements from laser scanning images of the optic nerve head [20]; employing an ensemble with special voting schemata for early melanoma diagnosis [31]; etc. Recently, Zhou and Jiang [41] have proposed the C4.5 Rule-PANE algorithm, which combines ensemble learning technique with C4.5Rule induction, and achieved strong generalization as well as good comprehensibility in medical tasks.

III. CO-FOREST

Let L and U denote the labeled set and unlabeled set respectively, which are drawn independently from the same data distribution. In co-training paradigm, two classifiers are firstly trained from L, and then each of them selects the most confident examples in U to label from its point of view, and the other classifier updates itself with these newly labeled examples. One of the most important aspect in co-training is how to estimate the confidence of a given unlabeled example. In standard co-training, the confidence estimation directly benefits from the two sufficient and redundant attribute subsets, where labeling confidence of a classifier could be regarded as its confidence for an unlabeled example. When the two sufficient and redundant attribute subsets do not exist, ten-fold cross validation is applied in each training iteration to estimate the confidence for the unlabeled data [17], in order not to bias its peer classifier with the unconfident examples. The ineffective confidence estimation greatly reduces the applicability of the extended co-training algorithm in many real-world applications such as computer-aided diagnosis.

However, if an ensemble of N classifiers, which is denoted by H^*, are used in co-training instead of two classifiers, the confidence could be estimated in an efficient way. When determining the most confidently labeled examples for a component classifier of the ensemble h_i ($i = 1, ..., N$), all other component classifiers in H^* except h_i are used. These component classifiers form a new ensemble, which is called the *concomitant ensemble* of h_i, denoted by H_i. Note that H_i differs from H^* only by the absence of h_i. Now the confidence for an unlabeled example can be simply estimated by the degree of agreements on the labeling, i.e. the number of classifiers that agree on the label assigned by H_i. By using this method, Co-Forest firstly trains an ensemble of classifiers on L and then refines each component classifier with unlabeled examples selected by its concomitant ensemble.

In detail, in each learning iteration of Co-Forest, the concomitant ensemble H_i examines each example in U. If the number of classifiers voting for a particular label exceeds a pre-set threshold θ, the unlabeled example along with the newly assigned label is then *copied* into the newly labeled set L'_i. Set $L \cup L'_i$ is used for the refinement of h_i in this iteration. Note that the unlabeled examples selected by H_i are not removed from U, so they might be selected again by other H_j ($j \neq i$) or the concomitant ensembles in the following iterations.

Since all the examples whose estimated confidence are above θ will be added to L'_i, the size of L'_i might be very large, even equal to the size of U in an extreme case. However, when the learned hypothesis has not fully captured the underlying distribution, especially in several initial iterations, using such a huge amount of automatically-labeled data might affect the performance of the learned hypothesis. Inspired by Nigam et al. [28], each unlabeled example is assigned a weight. Unlike the fixed weight used in [28], in our approach an example is weighted by the predictive confidence of a concomitant ensemble. On the one hand, the weighting of unlabeled example reduces the potential negative influence of the use of overwhelming amount of unlabeled data. On the other hand, it makes the algorithm insensitive to the parameter θ. Even if θ is small, this weighting strategy can limit the influence of the examples with low predictive confidence.

Furthermore, the use of an ensemble of classifiers here not only serves as a simple way to avoid utilizing complicated confidence estimation method, but also makes the labeling of the unlabeled data more accurate than a single classifier. However, although ensemble generalizes better than a single classifier, misclassification of unlabeled example is inevitable. So h_i receives noisy examples from time to time, which might bias the refinement of h_i. Fortunately, the following derivation inspired by Goldman and Zhou [17] shows that the negative influence caused by such noise could be compensated by augmenting the labeled set with sufficient amount of newly labeled examples under certain conditions.

According to Angluin and Laird [1], if the size of training data (m), the noise rate (η) and the hypothesis worst-case error rate (ϵ) satisfy the following relationship

$$m = \frac{c}{\epsilon^2(1-2\eta)^2} \tag{1}$$

where c is a constant, then the learned hypothesis h_i that minimizes the disagreement on a sequence of noisy training examples converges to the true hypothesis h^* with the probability equal to one.

By reforming (1), the following utility function is obtained.

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

In the t-th learning iteration, a component classifier h_i ($i = 1...N$) is supposed to refine itself on the union of original labeled set L with the size of m_0 and the newly labeled set $L'_{i,t}$ with the size of $m_{i,t}$, where $L'_{i,t}$ is determined and labeled by its concomitant ensemble H_i. Let the error rate of H_i on $L'_{i,t}$ be $\hat{e}_{i,t}$, and then the weighted number of examples being mislabeled by H_i in $L'_{i,t}$ is $\hat{e}_{i,t}W_{i,t}$, where $W_{i,t} = \sum_{j=0}^{m_{i,t}} w_{i,t,j}$ and $w_{i,t,j}$ is the predictive confidence of H_i on x_j in $L'_{i,t}$. To uniform the expressions, m_0 is rewritten as the weighted form W_0 where $W_0 = \sum_{j=0}^{m_0} 1$. In the augmented training set $L \cup L'_{i,t}$, the noisy examples consist of the noisy examples in L and the examples in $L'_{i,t}$ that are misclassified by the concomitant ensemble H_i. Thus the noise rate in $L \cup L'_{i,t}$ is estimated by

$$\eta_{i,t} = \frac{\eta_0 W_0 + \hat{e}_{i,t} W_{i,t}}{W_0 + W_{i,t}} \tag{3}$$

By replacing η and m in (2) with (3) and the weighted size of the augmented training set $(W_0 + W_{i,t})$ respectively, the utility of h_i in the t-th iteration takes the form of

$$u_{i,t} = (W_0 + W_{i,t}) \left(1 - 2\frac{\eta_0 W_0 + \hat{e}_{i,t} W_{i,t}}{W_0 + W_{i,t}}\right)^2 \tag{4}$$

Similarly, the utility of h_i in the $(t-1)$-th iteration is

$$u_{i,t-1} = (W_0 + W_{i,t-1}) \left(1 - 2\frac{\eta_0 W_0 + \hat{e}_{i,t-1} W_{i,t-1}}{W_0 + W_{i,t-1}}\right)^2 \tag{5}$$

According to (2), the utility u is inverse proportion to the squared worse-case error rate ϵ^2. Thus, to reduce the worst-case error rate of each classifier h_i, the utility of h_i should be increased in the learning iterations, i.e. $u_{i,t} > u_{i,t-1}$. Now assume that little noise exists in L and each component classifier h_i meets the requirement of weak classifier, i.e $\hat{e}_{i,t} < 0.5$. By comparing the right hand side of (4) and (5), $u_{i,t} > u_{i,t-1}$ holds when $W_{i,t} > W_{i,t-1}$ and $\hat{e}_{i,t}W_{i,t} < \hat{e}_{i,t-1}W_{i,t-1}$, which are further summarized by

$$\frac{\hat{e}_{i,t}}{\hat{e}_{i,t-1}} < \frac{W_{i,t-1}}{W_{i,t}} < 1 \tag{6}$$

According to (6), $\hat{e}_{i,t} < \hat{e}_{i,t-1}$ and $W_{i,t} > W_{i,t-1}$ should be satisfied at the same time. However, even if this requirement is met, $\hat{e}_{i,t}W_{i,t} < \hat{e}_{i,t-1}W_{i,t-1}$ might still be violated since $W_{i,t}$ might be much larger than $W_{i,t-1}$. To make (6) hold again in this case, $L'_{i,t}$ must be subsampled so that $W_{i,t}$ is less than $\frac{\hat{e}_{i,t-1}W_{i,t-1}}{\hat{e}_{i,t}}$.

Another important factor in co-training is how to produce the learned hypothesis with the refined classifiers, which is sometimes complicated and time-consuming [17]. Since an ensemble of classifiers is introduced to extend the co-training process, *majority voting*, which is widely used in ensemble learning, is employed to produce the final hypothesis.

Note that, when (6) holds, component classifiers are refined with unlabeled data, so the average error rate of component classifiers are expected to be reduced as the semi-supervised learning process proceeds. Nevertheless, the performance improvement of each component classifier does not necessarily lead to the performance improvement of the ensemble. According to Krogh and Vedelsby [24], an ensemble exhibits its generalization power when the average error rate of component classifiers is low and the diversity between component classifiers is high. To obtain a good performance of the ensemble, the diversity between component classifiers should be maintained when Co-Forest exploits the unlabeled data.

Unfortunately, the learning process of Co-Forest does hurt the diversity of classifiers. In each learning iteration, concomitant ensembles are used to select and label the unlabeled data for its corresponding classifiers. Since two concomitant ensembles H_i and H_j differs from each other only by two classifiers, i.e. h_i and h_j, the prediction made by H_i and H_j as well as the predictive confidence for each prediction could be quite similar, especially when the size of the concomitant ensembles is large. Thus, h_i and h_j will be similar in the next iteration after retraining themselves with the similar newly labeled sets. This degradation of diversity might counteract the error rate reduction of each component classifiers benefitting from the unlabeled data.

To maintain the diversity in the semi-supervised learning process, two strategies are employed. Firstly, a well-known ensemble method named *Random Forest* [7] is used to construct the ensemble in Co-Forest. Since Random Forest injects certain randomness in the tree learning process, any two trees in the Random Forest could still be diverse even if their training data are similar. Secondly, the diversity is further maintained when the concomitant ensembles select the unlabeled data to label. Specifically, not all the examples in U will be examined by concomitant ensembles. Instead, a subset of unlabeled examples with the total weight less than $\frac{\hat{e}_{i,t-1}W_{i,t-1}}{\hat{e}_{i,t}}$ is randomly selected from U. Then confident examples are further selected from the subset. Note that the subset not only offers diversity to some extent, but also acts as a pool to reduce the chance of being trapped into local minima, just like a similar strategy employed in [4].

TABLE I

PSEUDO-CODE DESCRIBING THE CO-FOREST ALGORITHM

Algorithm: Co-Forest
Input: the labeled set L, the unlabeled set U,
the confidence threshold θ, the number of random trees N
Process:
Construct a random forest consisting N random trees.
for $i \in \{1, ..., N\}$ **do**
  $\hat{e}_{i,0} \leftarrow 0.5$
  $W_{i,0} \leftarrow 0$
end for
$t \leftarrow 0$
Repeat until none of the trees in Random Forest changes
  $t \leftarrow t + 1$
  for $i \in \{1, ..., N\}$ **do**
    $\hat{e}_{i,t} \leftarrow EstimateError(H_i, L)$
    $L'_{i,t} \leftarrow \phi$
    if$(\hat{e}_{i,t} < \hat{e}_{i,t-1})$
      $U'_{i,t} \leftarrow SubSampled(U, \frac{\hat{e}_{i,t-1}W_{i,t-1}}{\hat{e}_{i,t}})$
      for each $x_u \in U'_{i,t}$ **do**
        if $(Confidence(H_i, x_u) > \theta)$
          $L'_{i,t} \leftarrow L'_{i,t} \cup \{(x_u, H_i(x_u))\}$
          $W_{i,t} \leftarrow W_{i,t} + Confidence(H_i, x_u)$
        end for
  end for
  for $i \in \{1, ..., N\}$ **do**
    if $(e_{i,t}W_{i,t} < e_{i,t-1}W_{i,t-1})$
      $h_i \leftarrow LearnRandomTree(L \cup L'_{i,t})$
  end for
end of Repeat
Output: $H^*(x) \leftarrow \arg\max_{y \in label} \sum_{i:h_i(x)=y} 1$

Table I shows the pseudo-code of Co-Forest. N random trees are firstly initiated from the training set bootstrap sampled from L for creating a Random Forest. Then, in each learning iterations each random tree is refined with the newly labeled examples selected by its concomitant ensemble under the conditions showing in (6), where the error rate $\hat{e}_{i,t}$ of concomitant ensemble H_i should be estimated accurately. Here, the error rate is estimated on the training data under the assumption that the incoming examples to be predicted come from the same distribution as that of training data. This method tends to under-estimate the error rate. Fortunately, since the Random Forest in Co-Forest is initiated through bootstrap sampling [14] on L, the *out-of-bag error* estimation [7] could be used at the first iteration to give a more accurate estimate of $\hat{e}_{i,t}$. This method reduces the chance of the trees in the Random Forest being biased when utilizing unlabeled data at the first iteration.

Note that by introducing ensemble method into the co-training process, Co-Forest requires neither the data described by the sufficient and redundant attribute subsets nor the use of two special supervised learning algorithms which frequently

use ten-fold cross validation to select the confident unlabeled examples to label and to produce the final hypothesis. Therefore, Co-Forest can be easily applied to many real-world applications such as computer-aided diagnosis. Moreover, Co-Forest extends tri-training [43] with more classifiers. These classifiers enable Co-Forest to exploit the power of ensemble in confidently selecting the unlabeled examples to label and producing the final hypothesis that generalizes quite well.

IV. Experiments

Ten data sets from UCI machine learning repository [2] are used in the experiments. Table II tabulates the detailed information of the experimental data sets. Among these data sets, three medical diagnosis data sets, namely *diabetes*, *hepatitis*, and *wpbc*, are further analyzed, respectively, to verify the effectiveness of proposed Co-Forest algorithm on medical diagnosis tasks. The *diabetes* data set is a collection of diabetes cases of Pima Indians from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 samples described by 8 continuous attributes. 268 samples are tested positive for diabetes and the other 500 samples are negative. The *hepatitis* data set consists of samples of 155 patients described by 19 attributes, i.e. 5 continuous attributes and 14 nominal ones. Among these patients, 32 patients died of hepatitis while the remaining ones survived. In the data set of *wpbc*, 33 continuous attributes are used to describe 198 samples belonging to two classes, i.e. whether the breast cancer would reoccur within 24 months.

TABLE II
Experimental Data Sets

Data set	# features	# instances	# classes
bupa	6	345	2
colic	22	368	2
diabetes	8	768	2
hepatitis	19	155	2
hypothyroid	25	3163	2
ionosphere	34	351	2
kr-vs-kp	36	3196	2
sonar	60	208	2
vote	16	435	2
wpbc	33	198	2

For each data set, 10-fold cross validation is employed for evaluation. In each fold, training data are randomly partitioned into labeled set L and unlabeled set U for a given *unlabel rate* (μ), which can be computed by the size of U over the size of $L \cup U$. For instance, if a training set contains 100 examples, splitting the training set according to unlabel rate 80% will produce a set with 20 labeled examples and a set with 80 unlabeled examples. In order to simulate different amount of unlabeled data, four different unlabel rates, i.e. 20%, 40%, 60% and 80%, are investigated here. Note that the class distributions in L and U are kept similar to that in the original data set.

As mentioned in Section III, the learning process of Co-Forest might hurt the diversity of the component classifiers when the size of ensemble is large. According to Zhou et al. [45], large size of ensemble does not necessarily lead to better performance of an ensemble. Thus, the ensemble size N in Co-Forest is not supposed to be too big. In the experiments, the value of N is set to 6. The other parameters of Random Forest adopt the default settings of the *RandomForest* package in WEKA [37]. The confident threshold θ is set to 0.75, i.e. an unlabeled example is regarded as being confidently labeled if more than $3/4$ trees agree on a certain label.

For comparison, the performance of two semi-supervised algorithms, i.e. *co-training* and *self-training*, are also evaluated. Since standard co-training [4] requires the sufficient and redundant attribute subsets, it could not be directly applied to the experimental data sets. Fortunately, previous work [27] indicates that under this circumstance co-training could still benefit from the unlabeled data in most of time by randomly splitting the attributes into two sets. Thus, the attributes in each data set are randomly split into two disjoint sets with almost equal size, just like what was done in [27], and then the co-training algorithm learns hypothesis from the transformed data set. The self-training algorithm [27] learns hypothesis from the labeled data and keeps on refining the hypothesis with the self-labeled data from the unlabeled set. Although the self-training algorithm has similar working style to the co-training algorithm, it has no requirement on the data sets. Note that the termination criteria in both standard co-training algorithm and self-training algorithm are different from that in Co-Forest. For fair comparison, the termination criteria of co-training and self-training are modified to that in Co-Forest. Random tree and Random Forest trained on L, denoted by *RTree* and *Forest* respectively, are used as the baselines for comparison. Here, random tree is the base classifier in Random Forest. The settings of random tree and Random Forest are kept the same as that in Co-Forest. These two baselines illustrate how well a Random Forest and one of its component can perform without further exploiting the unlabeled data, respectively. Moreover, SVM and AdaBoost [15] trained on L are also compared in the experiment, providing a reference to the performance achieved by some top classifiers without utilizing unlabeled data.

For each data set under a specific unlabel rate, 10-fold cross validation is repeated 10 times, and the results are averaged. Table III to Table VI tabulate the average error rates of the learned hypotheses under different unlabel rates. In the columns of the three semi-supervised learning algorithms, *initial* and *final* shows the average error rates of the hypotheses learned only with the labeled data and those further refined with the unlabeled data respectively. The performance improvement of the learned hypothesis from the unlabeled data is denoted by *improv.*, which can be computed by the reduction of error rates of the learned hypothesis over that of the hypothesis initially learned with the labeled data. Note that some of the data in the tables seem inconsistent due to truncation. The highest improvement under each unlabel rates on each data set has been boldfaced. Pairwise two-tailed t-test under the significance level 0.05 is applied to the experimental results, and the significant performance improvement is marked by a star. The row *avg.* in each table shows the average results over all the experimental data sets.

Moreover, classifiers are trained on $L \cup U$ provided with all ground-truth labels of the unlabeled data (i.e. the case when $\mu = 0\%$). Such data set is referred as *ideal training*

TABLE III

AVERAGE ERROR RATES OF THE COMPARED ALGORITHMS UNDER THE UNLABEL RATE OF 80%

Data set	RTree	Forest	SVM	AdaBoost	Self-Training			Co-Training			Co-Forest		
					initial	final	improv.	initial	final	improv.	initial	final	improv.
bupa	.396	.395	.420	.387	.396	.424	-7.1%*	.427	.443	-3.6%	.395	.384	**2.9%***
colic	.272	.208	.233	.230	.272	.278	-2.3%	.255	.285	-11.7%*	.208	.178	**14.5%***
diabetes	.321	.278	.261	.263	.321	.318	0.8%	.374	.356	4.8%*	.278	.261	**6.2%***
hepatitis	.231	.203	.186	.206	.231	.240	-4.2%	.246	.240	2.8%	.203	.180	**11.5%***
hypothyroid	.023	.018	.035	.014	.023	.023	-2.6%	.032	.038	-18.4%*	.018	.017	**6.6**%
ionosphere	.159	.129	.155	.156	.159	.191	-20.4%*	.179	.194	-8.6%	.129	.092	**28.7%***
kr-vs-kp	.100	.051	.055	.080	.100	.122	-22.2%*	.112	.123	-9.4%*	.051	.035	**32.2%***
sonar	.367	.312	.273	.306	.367	.388	-5.7%	.366	.398	-8.7%*	.312	.282	**9.7%***
vote	.088	.066	.056	.053	.088	.096	-9.2%	.104	.135	-30.1%*	.066	.056	**15.0%***
wpbc	.333	.328	.244	.304	.333	.373	-12.3%*	.353	.341	3.5%	.328	.279	**15.0%***
avg.	.229	.199	.192	.200	.229	.245	-8.5%	.245	.255	-7.9%	.199	.176	**14.2%**

TABLE IV

AVERAGE ERROR RATES OF THE COMPARED ALGORITHMS UNDER THE UNLABEL RATE OF 60%

Data set	RTree	Forest	SVM	AdaBoost	Self-Training			Co-Training			Co-Forest		
					initial	final	improv.	initial	final	improv.	initial	final	improv.
bupa	.396	.376	.422	.390	.396	.403	-1.7%	.411	.447	-8.7%*	.376	.364	**3.2%***
colic	.242	.189	.203	.199	.242	.258	-6.8%	.213	.240	-12.7%*	.189	.162	**14.2%***
diabetes	.318	.279	.243	.260	.318	.310	2.5%	.363	.365	-0.5%	.279	.264	**5.3%***
hepatitis	.239	.199	.180	.201	.239	.221	**7.3%**	.217	.227	-4.7%	.199	.186	6.7%
hypothyroid	.020	.014	.032	.011	.020	.021	-8.0%	.030	.034	-11.3%*	.014	.013	**8.8%***
ionosphere	.143	.104	.139	.135	.143	.142	0.8%	.151	.150	0.2%	.104	.079	**23.9%***
kr-vs-kp	.071	.033	.050	.084	.071	.075	-5.4%	.082	.079	3.7%	.033	.023	**30.8%***
sonar	.330	.282	.264	.297	.330	.347	-5.3%	.348	.341	1.9%	.282	.246	**12.8%***
vote	.076	.058	.048	.050	.076	.087	-14.2%*	.084	.120	-42.4%*	.058	.050	**13.9%***
wpbc	.312	.319	.231	.262	.312	.354	-13.6%*	.338	.329	2.6%	.319	.267	**16.1%***
avg.	.215	.185	.181	.189	.215	.222	-4.4%	.224	.233	-7.2%	.185	.165	**13.6%**

TABLE V

AVERAGE ERROR RATES OF THE COMPARED ALGORITHMS UNDER THE UNLABEL RATE OF 40%

Data set	RTree	Forest	SVM	AdaBoost	Self-Training			Co-Training			Co-Forest		
					initial	final	improv.	initial	final	improv.	initial	final	improv.
bupa	.379	.360	.419	.372	.379	.394	-4.0%	.431	.440	-2.0%	.360	.347	**3.7%**
colic	.242	.178	.187	.188	.242	.240	0.7%	.193	.209	-8.0%*	.178	.160	**10.1%***
diabetes	.304	.271	.231	.259	.304	.305	-0.2%	.362	.363	-0.4%	.271	.255	**5.9%***
hepatitis	.196	.184	.151	.186	.196	.215	-9.9%	.204	.209	-2.8%	.184	.163	**11.6%***
hypothyroid	.017	.012	.028	.010	.017	.018	-7.6%	.028	.032	-17.2%*	.012	.011	**9.8%***
ionosphere	.124	.093	.119	.124	.124	.128	-3.2%	.132	.142	-7.3%	.093	.075	**19.1%***
kr-vs-kp	.056	.026	.047	.081	.056	.058	-2.5%	.067	.071	-5.4%*	.026	.019	**27.7%***
sonar	.309	.269	.235	.287	.309	.310	-0.2%	.305	.310	-1.8%	.269	.224	**16.7%***
vote	.076	.058	.051	.044	.076	.066	**12.5%**	.085	.090	-6.3%	.058	.051	12.3%*
wpbc	.328	.305	.233	.264	.328	.348	-6.3%	.342	.315	7.9%	.305	.266	**12.5%***
avg.	.203	.176	.170	.182	.203	.208	-2.1%	.215	.218	-4.3%	.176	.157	**12.9%**

set thereinafter. Since all the examples are labeled, only the results of the baseline methods are shown in Table VII.

Table III to Table VI show that unlabeled data could be used to enhance the performance of the hypothesis learned only on the labeled data over different unlabel rates. Co-Forest achieves an overall 13.1% performance improvement. Under each unlabel rates, Co-Forest achieves significantly improvement on most of the data sets. The sign test applied on the results of t-test indicates that the improvement in the experiment is significant. It is also shown in the table that, after further exploiting the merit of unlabeled data, the hypothesis learned by Co-Forest reaches lower error rates than those learned by the baseline methods only on the labeled examples under all unlabel rates. Interestingly, when comparing the error rates of the baseline methods when $\mu = 0\%$ (i.e. the ideal training set) with those of Co-Forest, it could be observed that the hypothesis learned with certain amount of data unlabeled even outperforms those learned by the baseline methods with all the training data labeled. For example, when 80% data are unlabeled, Co-Forest, by exploiting the unlabeled examples, is able to reach an error rate comparable to that of AdaBoost using the ideal training set; when 60% examples are unlabeled, Co-Forest achieves comparable performance to SVM using the ideal training set, and outperforms AdaBoost using the ideal

TABLE VI
AVERAGE ERROR RATES OF THE COMPARED ALGORITHMS UNDER THE UNLABEL RATE OF 20%

Data set	RTree	Forest	SVM	AdaBoost	Self-Training			Co-Training			Co-Forest		
					initial	final	improv.	initial	final	improv.	initial	final	improv.
bupa	.378	.349	.421	.356	.378	.363	4.0%	.409	.421	-2.8%	.349	.331	**5.1%***
colic	.233	.170	.179	.181	.233	.222	4.8%	.193	.189	2.0%	.170	.158	**6.9%***
diabetes	.303	.268	.228	.259	.303	.306	-0.7%	.365	.353	**3.1%**	.268	.261	2.8%
hepatitis	.238	.183	.159	.192	.238	.216	**9.5%***	.200	.209	-4.5%	.183	.166	**9.5%***
hypothyroid	.016	.011	.027	.009	.016	.018	-7.4%	.028	.029	-3.8%	.011	.010	**10.6%***
ionosphere	.114	.085	.119	.124	.114	.119	-4.7%	.120	.131	-9.3%	.085	.069	**19.1%***
kr-vs-kp	.050	.019	.044	.089	.050	.048	5.2%	.062	.061	1.8%	.019	.014	**25.0%***
sonar	.276	.233	.227	.264	.276	.289	-4.8%	.294	.324	-10.0%*	.233	.201	**14.0%***
vote	.062	.053	.047	.045	.062	.073	-18.2%*	.081	.088	-8.5%	.053	.048	**10.8%***
wpbc	.312	.282	.230	.261	.312	.322	-3.3%	.340	.333	2.1%	.282	.250	**11.5%***
avg.	.198	.165	.168	.178	.198	.198	-1.6%	.209	.214	-3.0%	.165	.151	**11.5%**

training set.

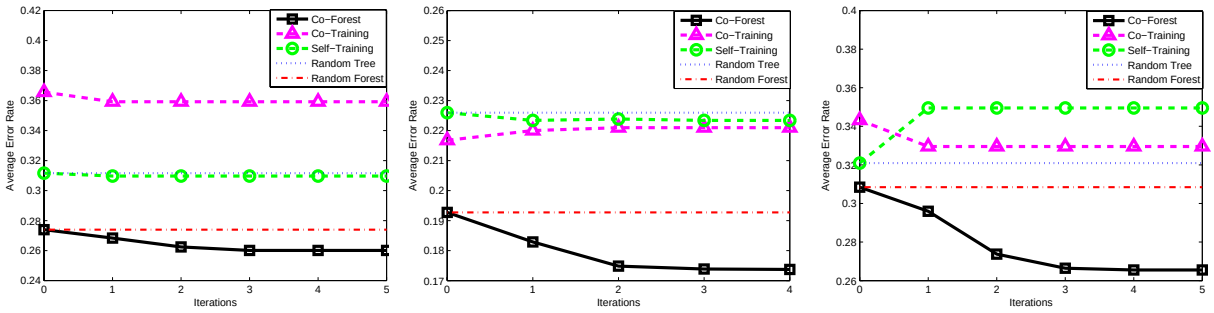
TABLE VII
AVERAGE ERROR RATES OF THE COMPARED ALGORITHMS UNDER THE UNLABEL RATE OF 0%

Data set	RTree	Forest	SVM	AdaBoost
bupa	.364	.347	.420	.358
colic	.214	.158	.170	.173
diabetes	.307	.268	.230	.259
hepatitis	.218	.183	.146	.202
hypothyroid	.017	.010	.026	.008
ionosphere	.111	.075	.121	.124
kr-vs-kp	.038	.017	.041	.093
sonar	.279	.231	.231	.272
vote	.071	.048	.044	.039
wpbc	.316	.258	.232	.250
avg.	.194	.160	.166	.178

While Co-Forest benefits from the unlabeled data, co-training and self-training fail to improve the performance of the learned hypotheses using the unlabeled data. Although the performance improvement is observed on some data sets under certain unlabel rates, in most cases the performance degrades after exploiting unlabeled data using co-training and self-training. By averaging on all the data sets and all the unlabel rates, the average error rate of co-training and self-training increases by 5.6% and 4.2%, respectively. Since the same termination criterion is employed in Co-Forest, co-training and self-training, the three algorithms differ from each other by the way of labeling unlabeled examples, which leads to different performance for utilizing the unlabeled data.

In self-training, there is only one classifier involved in the learning process, and thus the classifier has to provide the labels for unlabeled examples totally based on its current 'knowledge'. If the classifier is initially biased much, keeping on learning with the self-labeled examples makes the classifier overfit quickly, which leads to the performance degradation. The fewer the labeled training data, the more chance for the classifier to be biased, and hence the more chance for the performance degrades. This claim is confirmed by Table III to Table VI. By contrast, in Co-Forest each component classifier h_i is refined by the examples labeled by its concomitant ensemble H_i instead of itself, and thus, there is less chance for h_i to overfit. Moreover, since the label is assigned by an ensemble instead of a single classifier, h_i is more likely to receive correctly labeled examples than that in self-training. In co-training, the major reason accounting for the performance degradation is the violation of the requirement on sufficient and redundant views of data set. Since no experimental data set contains sufficient and redundant attribute sets, the original attribute set has to be randomly partitioned into two parts, which are not usually conditionally independent to each other given the class label. Thus, the classifiers trained on this two attribute set might behave similarly, such that the same unlabeled examples could be mislabeled by both classifiers. In the extreme case when all the examples mislabeled by the two classifiers are exactly the same, the effect of co-training degenerates to self-training. Moreover, since fewer attributes are used to train classifiers after the partitioning, the performance of learned classifiers could be worse than a classifier learned with the same amount of training data using original attribute sets. This claim is consistent with the tables, where initial error rates of co-training is much higher than the initial error rates of self-training. Due to the worse performance of component classifier, each component classifier is very likely to assign incorrect labels to the unlabeled data. Because of the second reason, co-training might even perform worse than self-training. This fact can also be observed in the tables. By contrast, Co-Forest works on original attribute sets and leverages the power of concomitant ensembles to provide the labeling for unlabeled examples.

In order to investigate the effectiveness of Co-Forest on medical diagnosis tasks, the performance of Co-Forest on *diabetes*, *hepatitis* and *wpbc* are further analyzed. It can be observed from the table that Co-Forest is able to enhance the performance of the learned hypothesis using unlabeled data under different unlabel rates. The average error rate over the four different unlabel rates reduces by 5.1% on *diabetes*, 9.8% on *hepatitis* and 13.8% on *wpbc*, respectively. By contrast, although co-training and self-training are able to benefit from the unlabeled data on the three tasks under certain unlabel rates, the improvement is quite limited and the error rates of the learned hypothesis are higher than Co-Forest. Besides, performance degradation can also be observed in the table, sometimes the degradation is rather drastic, e.g. the performance improves -13.6% when applying self-training on *wpbc*



(a) *diabetes* (b) *hepatitis* (c) *wpbc*

Fig. 1. Error rates averaged over different unlabel rates on experimental data sets

under unlabel rate 60%. It can be concluded that Co-Forest, which leverages the advantages of ensemble, is suitable for exploiting unlabeled data in conventional medical diagnosis tasks which have no sufficient and redundant attribute sets. The generalization ability of Co-Forest is better than the compared two semi-supervised learning algorithms.

To get an insight of the learning process of Co-Forest, the average error rates at each learning iteration are further averaged over the different unlabel rates on each data set. Note that unlike terminating the learning process at a fixed number of iterations (e.g. 10), the termination criterion allows Co-Forest to stop at any round. Fig. 1 gives the plots of the average error rates versus the learning iterations from the 0th round to the maximum round reached before the algorithm stops (e.g. the maximum round of Co-Forest on hepatitis is 4). The error rates at the termination are used as the error rates in the rounds after the termination in the figure. It could be observed from the figure that the line of Co-Forest is always below those of the other compared algorithms. The error rate of Co-Forest keeps on decreasing after utilizing unlabeled data, and converges quickly within just a few learning iterations. Since the maximum iterations required for convergence is quite small, the training of Co-Forest could be very fast. This advantage makes Co-Forest more appealing when exploiting unlabeled data in computer-aided diagnosis in that the systems can be updated very fast when new data, both labeled and unlabeled, are available.

Note that in previous experiments, N, the ensemble size, is fixed in Co-Forest. Different N values might affect the diversity of the ensemble, which might counteract the performance improvement acquired through exploiting the unlabeled data. Therefore, the performance of Co-Forest with different ensemble size N ($N = 3, ..., 10, 20, 50, 100$) is further investigated. Other experiment setups remain unchanged. The average performance improvements of Co-Forest are shown in Fig. 2. In the figure, Co-Forest achieves its highest improvement on all the three data sets when N is not too big. For instance, under the unlabel rate 80%, the ensemble size for highest improvement is 4 on *diabetes*, 4 on *hepatitis* and 6 on *wpbc*, respectively. When N is large enough (e.g. $N = 100$), the improvement becomes very small and negative improvement even appears, especially when $\mu = 80\%$. This observation confirms the claim in Section III that large size of the ensemble

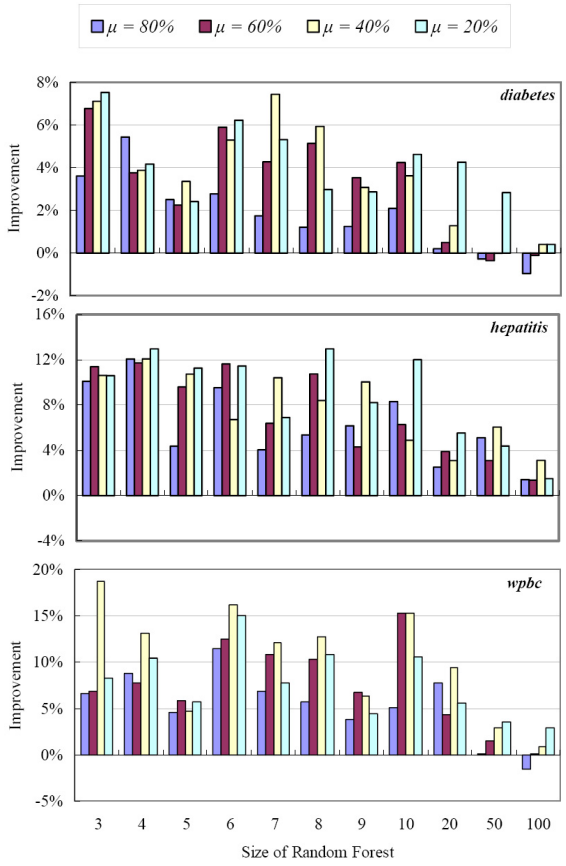
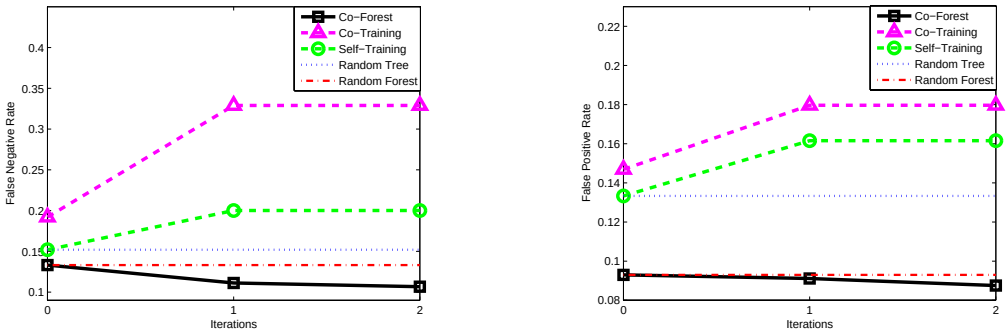


Fig. 2. Performance improvement over different ensemble sizes

leads to drastic decrease of diversity between component classifiers, and hence counteracts the benefits obtained by utilizing the unlabeled data. When the labeled training set is small, the initial diversity obtained by bootstrap sampling is limited. Consequently, the diversity may drop down rapidly as the learning proceeds, and the performance of the ensemble is severely humbled. This is why negative performance is usually observed when $\mu = 80\%$.

It is noteworthy that performance of Co-Forest varies on different data sets. For instance, its performance on *hepatitis* and *wpbc* are quite remarkable, but it performs not so impressive on *diabetes* as the other two data sets. This can be explained



(a) *False Negative Rate* (b) *False Positive Rate*

Fig. 3. The average false positive rates and false negative rates of compared algorithms

by the number of attributes in the data set. Since each tree in Random Forest is constructed by using the best attribute among several randomly selected attributes as the split at each internal node, the smaller the number of attributes, the more chance for some attributes to be selected together, hence more chance for the split to be the same. Thus, the trees in Random Forest trained on data set with fewer attributes could be less diverse than those trained on more attributes. Since the *diabetes* data set has only 8 attributes while *hepatitis* and *wpbc* have more attributes, it is possible for the improvement on *diabetes* to be less than those on *hepatitis* and *wpbc*. To solve this problem, new attributes might be generated according to method suggested by [7].

V. Application to Microcalcification Detection in Digital Mammograms

Breast cancer is the second leading cause of cancer death in woman, exceeded only by lung cancer. Since its pathogeny is unknown, breast cancer can hardly be prevented. The key for the survival of patients is the early detection of microcalcification clusters in digital mammograms, which is regarded as the aura of breast cancer.

The data set used here consists of 88 high-resolution digital mammograms collected from *Jiangsu Zhongda* Hospital, among which 20 images contain one or more microcalcification clusters marked by radiologists and the other 68 images are unmarked. Each digital image with 1914×2294 resolution and 12 bits pixel depth is fragmented into a set of 100×100 blocks. In each block, 5 features, i.e. *average density*, *density variance*, *energy variance*, *block activity* and *spectral entropy*, are extracted to form an example via the same method used in [22]. In the marked images, if there exists microcalcification in the block, the corresponding example is positive, otherwise it is a negative one. All the examples are left unlabeled if their corresponding blocks appear in the unmarked images. After removing the blocks of background, the data set comprises altogether 69 positive examples, 100 negative examples and additional 506 unlabeled examples. The goal of the learning system is to predict whether a block contains microcalcification clusters.

To evaluate the performance of Co-Forest on this microcalcification detection problem, five-fold cross validation is carried out, where the labeled data is partitioned in to 5 subsets with similar class distribution to that in the original labeled data. In each fold, classifiers are evaluated on one of the subset after being trained on the other four. The process of five-fold cross validation terminates after each subset has served as the test set exactly once, and the results are averaged over the 5 folds. In the experiment, the ensemble size of Co-Forest N is set to 6 and the confidence threshold θ is set to 0.75. For comparison, the co-training algorithm and the self-training algorithm are also evaluated here. Again, a random tree and a Random Forest trained only on the labeled data serve as the baselines for comparison. The parameters of the two baseline algorithms are kept the same as the corresponding ones in Co-Forest.

Since the early detection of microcalcification cluster leads to early cure of the disease, misclassifying the blocks with microcalcification as the normal ones reduces the chance for the survival of the patients. Thus, the *false negative rate*, which is computed by the ratio of the number of positive examples classified as negative by the learned hypothesis over the total number of examples are actually positive, becomes a major factor for evaluation of the algorithms. Moreover, since the doctors make their diagnosis according to the blocks detected by the system, misclassifying the normal blocks as lesions increases the burden on the doctors. Thus, *false positive rate*, which is computed by the ratio of the number of negative examples misclassified as positive over the number of examples classified as positive. Five-fold cross validation is repeated 10 times. Both the average false negative rates and the average false positive rates of all the algorithms versus the number iterations are plotted in Fig. 3.

Fig. 3 shows that Co-Forest benefits from the unlabeled data, and the learned hypothesis outperforms those learned by other compared algorithms. After two learning iterations, the average false negative rate decreases from the 0.133 to 0.107, which is lower than the two baselines. It is quite impressive that the average false negative rate of the learned hypothesis reduces by 20.0%. Meanwhile, the average false positive rate of the hypothesis learned by Co-Forest reduces by 5.8%. The reduction of both false negative rate and false positive rate suggests that without classifying more normal blocks as the positive ones, Co-Forest is able to use the unlabeled data to

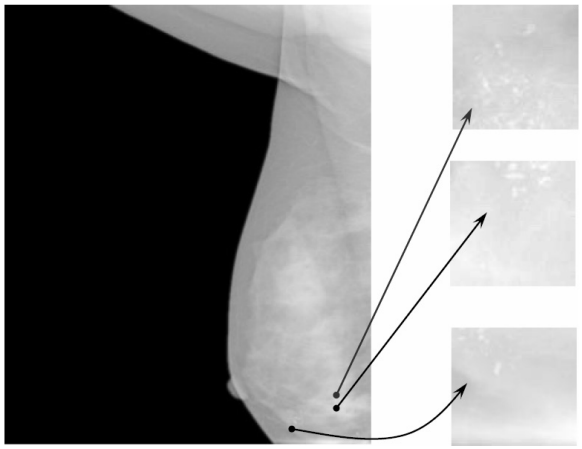


Fig. 4. The mammogram and the blocks with microcalcification detected after using unlabeled data

increase the chance of detecting microcalcification clusters in mammograms.

By contrast, while Co-Forest improves the performance with the unlabeled data, co-training and self-training fail to solve the microcalcification cluster detection problem. The performance of both co-training and self-training degenerates respectively after the unlabeled data are used to refine the learned hypothesis. The average false positive rate of co-training and self-training reduces by -22.3% and -21.3%, respectively, and the false negative rate of them even reduces by -71.2% and -31.7%. As shown in the figure, the curves of co-training and self-training are much higher than the curve of Co-Forest. Note that co-training exhibits very poor performance when handling this task. As explained in Section IV, the reason is that the microcalcification cluster detection problem does not contain sufficient and redundant attribute sets. Co-training has to work on randomly partitioned attribute sets. Since there are only 5 features, after the partitioning, it is difficult to discriminate the positive and negative examples using the 2 or 3 features in each view. Thus, each co-trained classifier tends to receive many unlabeled examples with incorrect labels from its peer classifier. As learning proceeds, the performance degrades quickly. Therefore, it is concluded that Co-Forest is a better solution to the microcalcification cluster detection.

To illustrate how Co-Forest benefits from the unlabeled data, a mammogram reduced to 292 $\times$ 350 resolution and three 100 $\times$ 100 blocks with microcalcification clusters in the mammogram are shown in Fig. 4, where the positions of the blocks in the mammogram have been marked. In the three selected blocks, the microcalcification clusters are neglected firstly by the hypothesis learned from only the labeled data, and then correctly detected after exploiting the unlabeled data with Co-Forest. Note that some of the microcalcification clusters are not very apparent in these blocks, which means that unlabeled data help the learning system focus on those unapparent areas sharing something in common with the areas that has been correctly identified as the lesions.

In summary, unlabeled data are beneficial in microcalcification detection. While co-training and self-training are ineffective for this task, the Co-Forest algorithm is able to enhance the performance of the learned hypothesis by exploiting unlabeled data in an effective and efficient way. Now Co-Forest is being implemented in the CabCD (Computer-aided breast Cancer Diagnosis) System by *Jiangsu Zhongda* Hospital.

VI. Conclusion

In computer-aided medical diagnosis, diagnosing the samples for training a well-performed CAD system places heavy burden on medical experts. Such burden could be released if the learning algorithm could use unlabeled data to help learning. In this paper, the Co-Forest algorithm is proposed, which can use undiagnosed samples to boost the performance of the system trained from the diagnosed samples. By extending the co-training paradigm, it exploits the power of Random Forest, a well-known ensemble method, to tackle the problem of selecting confident undiagnosed samples to label and producing the final hypothesis. Experiments on UCI data sets verify the effectiveness of Co-Forest. Case studies on three medical data sets and a successful application to microcalcification cluster detection for breast cancer diagnosis show that the undiagnosed samples are beneficial in building computer-aided diagnosis systems and Co-Forest is able to enhance the performance of the hypothesis learned simply on a small amount of diagnosed samples by exploiting the undiagnosed samples.

Since Co-Forest tends to under-estimate the error rates of the concomitant ensembles, finding an efficient method to properly estimate the error rates of these ensembles will be done in future, which is anticipated to make Co-Forest perform better. Another interesting future work is to enhance the performance of Co-Forest by incorporating *Query by Committee* [33], an active learning [9] mechanism, such that more helpful information can be provided by the diagnosis from medical experts on certain undiagnosed samples. Such a idea of combining semi-supervised learning with active learning in co-training paradigm has been applied for content based image retrieval [40]. Furthermore, it is noteworthy that the diversity between component classifiers is maintained based on the randomness provided by Random Forest. This places constraints to the base learner of Co-Forest and the scale of ensemble. In future, exploring a method to maintain the diversity of component classifiers in any ensembles during the semi-supervised learning process will extend the idea of Co-Forest to more general cases, such that it can be applied in more practical applications.

Acknowledgment

The comments and suggestions from the anonymous reviewers greatly improved this paper. The authors wish to thank their collaborators at the Jiangsu Zhongda Hospital for providing the high-resolution digital mammograms and their collaboration in developing the diagnosis system.

References

[1] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol.2, no.4, pp.343-370, 1988.

[2] C. Blake, E. Keogh, and C.J. Merz, "UCI repository of machine learning databases" [http://www.ics.uci.edu/∼mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[3] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, 2001, pp.19-26.

[4] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, WI, 1998, pp. 92-100.

[5] L. Breiman, "Bagging predictors," *Machine Learning*, vol.24, no.2, pp.123-140, 1996.

[6] L. Breiman, "Bias, variance, and arcing classifiers," *Technical Report*, University of California, Berkeley, CA, 1996.

[7] L. Breiman, "Random forests," *Machine Learning*, vol.45, no.1, pp.5-32, 2001.

[8] O.Chappelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006

[9] D.A. Cohn, Z. Ghahramani, and M.I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol.4, pp.129-145, 1996.

[10] P. Cunningham, J. Carney, and S. Jacob, "Stability problems with artificial neural networks and the ensemble solution," *Artificial Intelligence in Medicine*, vol. 20, no.3, pp.217-225, 2000.

[11] S. Dasgupta, M. Littman, and D. McAllester, "PAC generalization bounds for co-training," in T.G. Dietterich, S. Becker, and Z. Ghahramani, Eds., *Advances in Neural Information Processing Systems 14*, Cambridge, MA: MIT Press, pp.375-382, 2002.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol.39, no.1, pp.1-38, 1977.

[13] T.G. Dietterich, "Ensemble learning," in *The Handbook of Brain Theory and Neural Networks*, 2nd edition, M.A. Arbib, Ed., Cambridge, MA: MIT Press, 2002.

[14] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, New York: Chapman & Hall, 1993.

[15] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," in *Proceedings of the 2nd European Conference on Computational Learning Theory*, Barcelona, Spain, 1995, pp.23-37.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol.28, no.2, pp.337-407, 2000.

[17] S. Goldman and Y. Zhou, "Enhancing supervised learning with unlabeled data," in *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA, 2000, pp.327-334.

[18] L. Hansen and P. Salamon, "Neural network ensemble," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.12, no.10, pp.993-1001, 1990.

[19] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no.8, pp.832-844, 1998.

[20] T. Hothorn and B. Lausen, "Bagging tree classifiers for laser scanning images: a data- and simulation-based strategy," *Artificial Intelligence in Medicine*, vol.27, no.1, pp.65-79, 2003.

[21] R. Hwa, M. Osborne, A. Sarkar, and M. Steedman,"Corrected co-training for statistical parsers," in *Working Notes of the ICML'03 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.

[22] X. Jia, Z. Wang, S. Chen, N. Li, and Z.-H. Zhou, "Fast screen out true negative regions for microcalcification detection in digital mammograms," *Technical Report*, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 2005.

[23] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, 1999, pp. 200-209.

[24] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in G. Tesauro, D.S. Touretzky, and T.K. Leen, Eds., *Advances in Neural Information Processing Systems 7*, Cambridge, MA: MIT Press, 1995, pp. 231-238.

[25] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 626-633.

[26] D.J. Miller and H.S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled Data," in M. Mozer, M.I. Jordan, and T. Petsche, Eds., *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, 1997, pp. 571-577.

[27] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the 9th ACM International Conference on Information and Knowledge Management*, McLean, VA, 2000, pp. 86-93.

[28] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol.39, no.2-3, pp.103-134, 2000.

[29] D. Pierce and C. Cardie, "Limitations of co-training for natural language learning from large data sets," in *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA, 2001, pp. 1-9.

[30] A. Sarkar, "Applying co-training methods to statistical parsing," in *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, 2001, pp. 95-102.

[31] A. Sboner, C. Eccher, E. Blanzieri, P. Bauer, M. Cristofolini, G. Zuniani, and S. Forti, "A multiple classifier system for early melanoma diagnosis," *Artificial Intelligence in Medicine*, vol.27, no.1, pp.29-44, 2003.

[32] M. Seeger, "Learning with labeled and unlabeled data," *Technical Report*, University of Edinburgh, Edinburgh, Scotland, 2001.

[33] H. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the 5th ACM Workshop on Computational Learning Theory*, Pittsburgh, PA, 1992, pp. 287-294.

[34] B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol.32, no.5, pp. 1087-1095, 1994.

[35] A. Sharkey, N. Sharkey, and S. Cross, "Adapting an ensemble approach for the diagnosis of breast cancer," in *Proceedings of the 6th International Conference on Artificial Neural Networks*, Skövd, Sweden, 1998, pp. 281-286.

[36] M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim, "Bootstrapping statistical parsers from small data sets," in *Proceedings of the 10th Conference on the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003, pp. 331-338.

[37] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco: Morgan Kaufmann, 2000.

[38] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B Schölkopf, "Learning with local and global consistency", in S. Thrun, L. K. Saul, B. Schölkopf, Eds., *Advances in Neural Information Processing Systems 16*, Cambridge, MA: MIT Press, 2003, pp.1633-1640.

[39] Y. Zhou and S. Goldman, "Democratic co-learning", in *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, FL, 2004, pp. 594-602.

[40] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai, "Enhancing relevance feedback in image retrieval using unlabeled data," *ACM Transactions on Information Systems*, vol. 24, no. 2, pp. 219-244, 2006.

[41] Z.-H. Zhou and Y. Jiang. "Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble", *IEEE Transactions on Information Technology in Biomedicine*, vol.7, no.1, pp.37-42, 2003.

[42] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen. "Lung cancer cell identification based on artificial neural network ensembles", *Artificial Intelligence in Medicine*, vol.24, no.1, pp.25-36, 2002.

[43] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol.17, no.11, pp.1529-1541, 2005.

[44] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training," in *Proceedings of 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005, pp.908-913.

[45] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol.137, no.1-2, pp.239-263, 2002.

[46] X. Zhu, "Semi-supervised learning literature survey", *Techincal Report 1530*, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, 2005.

[47] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning*, Washinton DC, 2003, pp. 912-919.

Ming Li received the BSc degree in computer science from Nanjing University, China, in 2003. Currently, he is a PhD candidate at the Department of Computer Science & Technology of Nanjing University, and is a member of the LAMDA Group. His research interests mainly include machine learning and data mining, especially in learning with labeled and unlabeled examples. He has won a number of awards including the Microsoft Fellowship Award (2005), the HP Chinese Excellent Student Scholarship (2005), the Outstanding Graduate Student of Nanjing University (2006), etc. He won the PAKDD'06 Data Mining Competition Open Category Champion with other LAMDA members.

Zhi-Hua Zhou (S'00-M'01-SM'06) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as a lecturer in 2001, and is a professor and head of the LAMDA group at present. His research interests are in artificial intelligence, machine learning, data mining, information retrieval, pattern recognition, evolutionary computation, and neural computation. In these areas he has published over 40 papers in leading international journals or conference proceedings. He has won various awards/honors including the National Science & Technology Award for Young Scholars of China (2006), the Award of National Science Fund for Distinguished Young Scholars of China (2003), the National Excellent Doctoral Dissertation Award of China (2003), and the Microsoft Young Professorship Award (2006). He is on the editorial boards of *Knowledge and Information Systems*, *Artificial Intelligence in Medicine*, the *International Journal of Data Warehousing and Mining*, the *Journal of Computer Science & Technology* and the *Journal of Software*, and was guest editor/co-editor of journals including *ACM/Springer Multimedia Systems* and *The Computer Journal*. He served as the program committee chair for PAKDD'07, vice chair for ICDM'06, PRICAI'06, etc., program committee member for various international conferences including ICML, ECML, SIGKDD, ICDM, and chaired a number of native conferences. He is a senior member of China Computer Federation (CCF) and the vice chair of the CCF Artificial Intelligence & Pattern Recognition Society, an executive committee member of Chinese Association of Artificial Intelligence (CAAI) and the chair of the CAAI Machine Learning Society, a member of AAAI and ACM, and a senior member of IEEE and IEEE Computer Society.