Breast Cancer Survivability Prediction with Labeled, Unlabeled, and Pseudo-Labeled Patient Data

Juhyeon Kim, Hyunjung Shin*

Department of Industrial Engineering, Ajou University,
Wonchun-dong, Yeongtong-gu, Suwon 443-749, South Korea
{juhyeon, shin}@ajou.ac.kr

Abstract

Prognostic study on breast cancer survivability has been aided by machine learning algorithms which provide prediction on the survival of a particular patient on the basis of historical patient data. A labeled patient record however, is not easy to collect. It takes at least five years to label a patient record as "survived" or "not survived": meanwhile, unguided trials on numerous types of oncology-therapy cost highly. Moreover, it requires confidentiality agreements from both doctors and patients to obtain a labeled patient record. The difficulties in collection of labeled patient data have drawn researchers' attention to Semi-Supervised Learning (SSL), one of the most recent machine learning algorithms, since it is capable of utilizing unlabeled patient data as well which relatively much easier to collect, and therefore is regarded as a pertinent algorithm to circumvent the difficulties. However, the fact is yet valid even on SSL that more labeled data lead to better prediction. To make up for insufficiency of labeled patient data, one may consider an idea of tagging virtual labels to unlabeled patient data, namely "pseudo-labels", and using them as if they are labeled. The proposed algorithm, "SSL Co-training", implements the idea based on SSL. SSL Co-training was tested on the surveillance, epidemiology, and end results database for breast cancer (SEER) and achieved avg. 76% accuracy and avg. 0.81 AUC.

1. Introduction

Breast cancer is the most common type of cancer and the second leading cause of cancer deaths in women [1] [2]. The major clinical problem associated with breast cancer is to predict the outcome (survival or death) after the onset of therapeutically resistant disseminated disease. In many cases, by the time the primary tumor is diagnosed, clinically evident metastases have already occurred. In general, treatments such as chemotherapy, hormone therapy, or a combination are considered to reduce the spread of breast cancer by decreasing the distant metastases, by one-third. Therefore, the ability to predict disease outcomes more accurately would allow physicians to make informed decisions on the potential necessity of adjuvant treatment. This could also lead to the development of individually tailored treatments to maximize treatment efficiency [3] [4]. There are three predictive foci of cancer prognosis: (1) prediction of cancer susceptibility (risk assessment), (1) prediction of cancer recurrence (redevelopment of cancer after resolution), and (3) prediction of cancer survivability. In the third case, research focuses on predicting the outcome in terms of life expectancy, survivability, progression, or tumor-drug sensitivity after the diagnosis of the disease. In this study, we focus on the survivability prediction which involves the use of methods and techniques for predicting the survival of a particular patient on the basis of historical data [5]. In general, "survival" can be defined as the patient remaining alive for a specified period after the diagnosis of the disease. If the patient is still living for 1,825 days (5 years) after the date of diagnosis, then the patient is considered to have survived [6]. Note that the prediction on survivability is predominately used for the analysis where the interest is in observing time to death of a patient, but in this study, it is dealt with as a classification problem that predicts whether the patient belongs to the group of those who survived after a specified period.

Research on breast cancer with data mining or machine learning methods has led to improved treatments in the form of less-invasive predictive medicine. In [20] the authors conducted a wide-ranging investigation of different machine learning methods, discussing issues related to the types of data incorporated and the performance of these techniques in breast cancer prediction and prognosis. This review provides detailed explanations leading to first-rate research guidelines for the application of machine learning methods to cancer

[*] Corresponding author: Hyunjung (Helen) Shin, shin@ajou.ac.kr

prognosis. The authors of [5] used two popular data mining algorithms, artificial neural networks and decision trees, together with a common statistical method, logistic regression, to develop prediction models for breast cancer survivability. The decision tree turned out to be the best predictor. An improvement in the results of decision trees for the prognosis of breast cancer survivability is described in [4]. The authors propose a hybrid prognostic scheme based on weighted fuzzy decision trees. This hybrid scheme is an effective alternative to crisp classifiers that are applied independently. It analyzes the hybridization of accuracy and interpretability in terms of fuzzy logic and decision trees. In [21], the authors carried out data pre-processing using the RELIEF attribute selection and then used the Modest AdaBoost algorithm to predict breast cancer survivability. They used the Srinagarind hospital database. The results showed that Modest AdaBoost performs better than Real and Gentle AdaBoost. They then [22] proposed a hybrid scheme to generate a high-quality data set to develop improved breast cancer survival models.

To build such predictive models, a large quantity of breast cancer patient data is required. In machine learning or data mining domain, the types of data is categorized into "labeled (feature/label pairs)" and "unlabeled (features without labels)". With the patient data for breast cancer survivability, the label means the information tagged as "survived" if the patient survived after a specified period or "not survived" if he/she could not make it. Accumulating a substantial quantity of labeled data is time-consuming, costly, and requires confidentiality agreements. In general, the collection of labeled survival data requires at least five years [5-6]. Moreover, oncologist consultation fees must be paid to confirm survivability. Furthermore, doctors and patients seldom reveal their information. Now, the subject of inquiry is that in order to acquire the survival data whether is it worthy to wait for five years, pay significant amount of fee and exert a great deal efforts to convince patients to disclose their personal medical data? On the other hand, unlabeled data can be collected with much less efforts. In survival analysis, censored data are abundant because there are many cases that patient data have not been updated along time, and hence unlabeled. Then, an economical solution may be to utilize a large quantity of unlabeled data when building a predictive model. This becomes available with semi-supervised learning (SSL) algorithm that has recently emerged in the machine learning domain. SSL is an appealing method in areas where labeled data is hard to collect. It has been used in areas such as text classification [10], text chunking [11], document clustering [8], time-series classification [12], gene expression data classification [13-14], visual classification [15], question-answering task for ranking candidate sentences [16], and webpage classification [17]. As those examples in other domains, SSL would be a good idea since it is able to use the censored data to either modify or reprioritize the predictions on survivability obtained from labeled patient data alone. A good example of SSL employed for the prognosis of breast cancer survivability can be found in [41], where the successful implementation of SSL offered predictability of survival outcomes with reasonable accuracy and stability, relieving oncologists of the burden of collection for labeled patient data.

Although SSL is capable of utilizing unlabeled patient data, the prediction accuracy of SSL increases when the amount of labeled patient data increases like most algorithms in machine learning. To take into account the aforementioned difficulties in collection of labeled patient data, an idea to obtain more labeled data is to generate labels for unlabeled data and use them as if they are labeled. One may name them as "pseudo-labeled" data. This is the motivation of our study. The model proposed is called SSL Co-training. The model is based on SSL, and is composed of more than two member models in order to generate pseudo-labels. Unlabeled data become pseudo-labeled when agreement on labeling is reached among the member models. This process is repeated until no more agreement is obtained. The raised prediction accuracy for breast cancer survivability by labeled, unlabeled, and pseudo-labeled patient data would allow medical oncologists to perform more pertinent treatment for the cancer patients.

The remainder of the paper is organized as follows. Section 2 introduces SSL which is the base algorithm of the proposed Co-training algorithm. Section 3, the proposed algorithm SSL co-training is explained in length. Section 4 provides the experimental results for the comparative analysis between the proposed algorithm and the up-to-date machine learning models such as support vector machines (SVM), artificial neural networks (ANN), graph-based SSL. We use the surveillance, epidemiology, and end results (SEER) cancer incidence database, which is the most comprehensive source of information on cancer incidence and survival in the United States [18]. Section 5 presents the conclusions.

2. Semi-Supervised Learning

In many real world classification problems, the number of class-labeled data points is small because they are often difficult, expensive, or time-consuming to acquire, requiring qualified human annotators, as described in [29, 34, 37]. On the other hand, unlabeled data can easily be gathered and can provide valuable information for learning, as stated in [30]. However, traditional classification algorithms such as supervised-learning algorithms use only labeled data; therefore, they encounter difficulties when only a few labeled data are given. SSL uses both labeled and unlabeled data to improve on the performance of supervised learning; see as [30-31]. In SSL, the classification function is trained with a small set of labeled data $\{L = \{(x_i, y_i)_{i=1}^{n_l}\}$ and a large set of unlabeled data $U = \{(x_j)_{j=n_l+1}^{n}\}$, where $y = \pm 1$ indicates the labels. The total number of data points is $n = n_l + n_u$ [32]. There are several types of SSL algorithms, and graph-based SSL is used in our study. In graph-based SSL, a weighted graph is constructed in which the nodes represent the labeled and unlabeled data points and the edges reflect the similarity between data points. According to [33], graph-based SSL methods are nonparametric, discriminative, and transductive in nature. They assume label smoothness over the graph. This assumption states that if two data points are coupled by a path of high density (e.g., it is more likely that both belong to same group or cluster), then their outputs are likely to be close, whereas if they are separated by a low-density region then their outputs need not be close [31]. There are many graph-based SSL algorithms, e.g., mincut, Gaussian random fields and harmonic functions, local and global consistency, Tikhonov regularization, manifold regularization, graph kernels from the Laplacian spectrum, and tree-based Bayes [33]. There are many differences in the technical details, but in all these methods the labeled nodes are set to the labels $y_1 \in \{-1, +1\}$, the unlabeled nodes are set to zero ($y_u = 0$), and the pairwise relationships between nodes are represented via a similarity matrix [34]. Figure 1 depicts a graph with eight data points linked by similarity between them.

$$w_{ij} = \left\{ \begin{matrix} \exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{\alpha^2}\right) & if \quad i \sim j \\ 0 & otherwise \end{matrix} \right\} \tag{1}$$

The similarity between the two nodes x_i and x_j is represented via w_{ij} in a weight matrix W. Now, a label can propagate from (labeled) node x_i to node (unlabeled) node x_j only when the value of w_{ij} is large. The value of w_{ij} can be measured using the Gaussian function [31]:

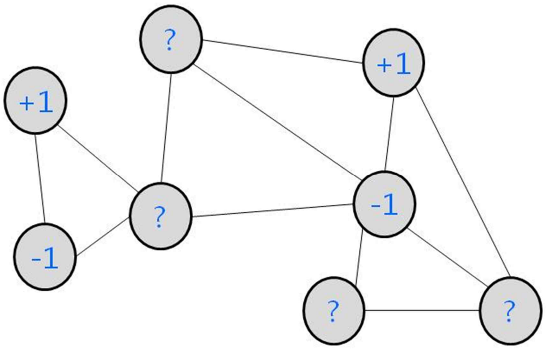


Figure 1. Graph-based SSL: Labeled nodes are represented by +1 (survivedl) and -1 (not survived), and unlabeled nodes are represented by ?.

In Eq. (1), i ~ j indicates that an edge (link) can be constructed between nodes x_i and x_j by the k-nearest neighbors algorithm, where k is a user-specified hyperparameter. The algorithm will output an n-dimensional real-valued vector $f = [f_l^T f_u^T]^T = (f_1, \ldots, f_l, f_{l+1}, \ldots, f_{n=l+u})^T$ that can generate a threshold value to carry out label predictions on $(f_1, \ldots, f_n)$ as a result of the learning. There are two assumptions: a loss function (f_i should be close to the given label of y_i in labeled nodes) and label smoothness (overall, f_i should not be too different from the f_i of the neighboring nodes). These assumptions are reflected in the value of f by minimizing the following quadratic function [29, 34, 35, 36]:

$$\min_f (f - y)^T (f - y) + \mu f^T L f, \tag{2}$$

where $y = (y_1, \ldots, y_l, 0, \ldots, 0)^T$ and the matrix L, called the graph Laplacian matrix, is defined as $L = D - W$ where $D = diag\ (d_i), d_i = \sum_i w_{ij}$. The parameter μ trades off loss and smoothness. The solution of this problem becomes

$$f = (I + \mu L)^{-1} y. \tag{3}$$

3. Proposed Method: Semi-Supervised Co-Training

For prediction of cancer survivability, SSL would be a good candidate to employ as a predictive model, particularly when an available dataset for model learning is abundant in unlabeled patient cases but lack of labeled ones. As many of other machine learning algorithms, however, it is also applied to SSL as well that more labeled data lead to better performance. A trick to obtain more labeled data is to assign labels for unlabeled data, namely, "pseudo-labels" and then use them for model learning as if they are labeled. The proposed model is about how to generate pseudo-labels, which eventually raises the performance of SSL. The model consists of multiple member models since pseudo-labels are determined based on agreement among the members. Therefore it is named as SSL Co-training. In this section, SSL Co-training is described limiting the number of members to two for the sake of simplicity.

The proposed algorithm is presented in Figure 2. Let L and U denote the sets of labeled and unlabeled datasets, respectively. And assume that two member models, F_1 and F_2, are given (more concretely, two SSL classifiers) and they are independent. In the beginning of the algorithm, each of the two classifiers is trained on L and U following the objective function in (2) as an ordinary SSL classifier. After training, both classifiers produce two sets of prediction scores for U according to (3). Let denote them as f_1 and f_2, respectively. The values of f_1 are continuous, thus discretization is required to make binary labels for U. A simple rule, by setting the midpoint of f_1 as the cutoff value m_1, provides labels to all of the unlabeled data: $y_u^1 = 1$ if f_1 is larger than m_1, $y_u^1 = -1$ otherwise. For the classifier F_2, y_u^2 is similarly obtained from the prediction score f_2 and its midpoint of m_2. And now the labels by F_1 may concordant or conflict with those of F_2. For unlabeled data points in U, the algorithm assigns pseudo-labels y_u only when all of the members agree on labeling since it gives higher confidence on the newly made labels. A unlabeled data point takes the value of its pseudo-label y_u either from F_1 or from F_2 when $y_u^1 = y_u^2$, or it remains unlabeled. The unlabeled data points that failed to obtain pseudo-labels are called "boosted samples." In the next iteration, the unlabeled data points with pseudo-labels are added to the labeled dataset L, whereas the boosted samples still belong to the unlabeled dataset U. As the iteration proceeds, therefore, the size of L increases while that of U decreases. The iteration stops if the size of U (the number of boosted samples) stops decreasing. Figure 3(a) depicts the decreasing pattern of the number of boosted samples while iteration goes. And Figure 3(b) shows the increasing pattern of model performance thanks to increasing size of labeled data points (note that the performances of the two member classifiers also increase). A toy example in Figure 4 will be helpful to understand the proposed algorithm.

The way of member composition for SSL Co-training can be diverse. First, the number of members is not limited, thus can be multiple. Second, different member models can be built from different data sources or different model parameters. In the current study, the two member models, F_1 and F_2, were built by splitting a dataset into two sub-datasets. The split is conducted so that the two sub-sets are maximally uncorrelated, i.e. the attributes in one set are uncorrelated to those in the other set.

```
$L$: labeled $\{(x_l, y_l)\}$ $y_l \in \{-1, 1\}$
$U$: unlabled $\{x\}$ $y_u \in \{0\}$
$F_1$: SSL classifier built on data set of $V_1$
$F_2$: SSL classifier built on data set of $V_2$
$f_1$: set of value predicted from the $f_1$
$f_2$: set of value predicted from the $f_2$
do
    Training $F_1$ and $F_2$
    $m_1 = $ midpoint of $f_1$
    $m_2 = $ midpoint of $f_2$
    $y_u^1 = \begin{cases} 1 & \text{if } f_u^1(x_u) > m_1 \\ -1 & \text{elsewhere} \end{cases}$
    $y_u^2 = \begin{cases} 1 & \text{if } f_u^2(x_u) > m_2 \\ -1 & \text{elsewhere} \end{cases}$
Labeling
    $y_i \leftarrow \frac{y_u^1 + y_u^2}{2}$ ($i = $ index set of $U$)
    if $y_{u_i}^1 + y_{u_i}^2$ (Agreement) $i = $ index of $u$
        $U \leftarrow U \setminus \{(x_i, y_i)\}$
        $L \leftarrow L \cup \{(x_i, y_i)\}$
    else (disagreement)
        $(x_i, y_i)$ remain as boosted samples
        (unlabeled data points for the next iteration)
while (if decrease in number of boosted samples)
```

Figure 2. SSL Co-training algorithm.

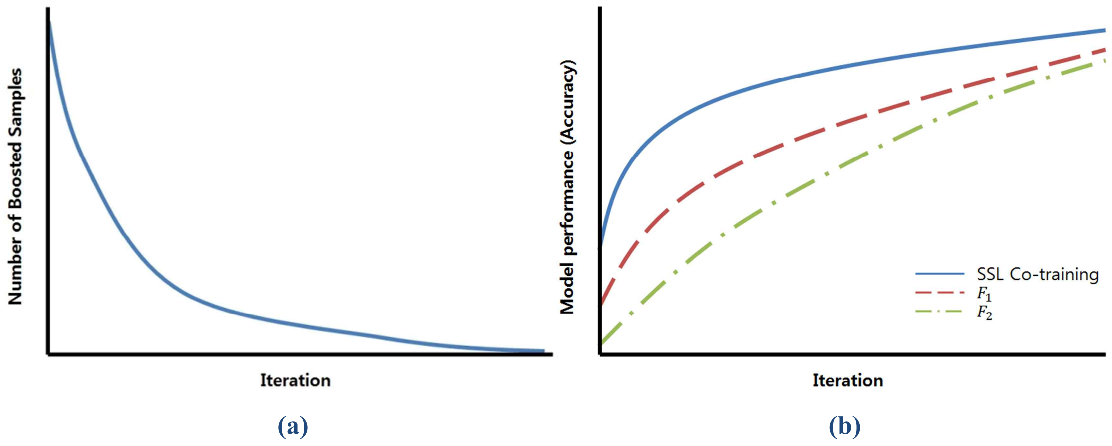


Figure 3. Patterns of (a) the number of boosted samples and (b) model performance during iterations of SSL Co-training

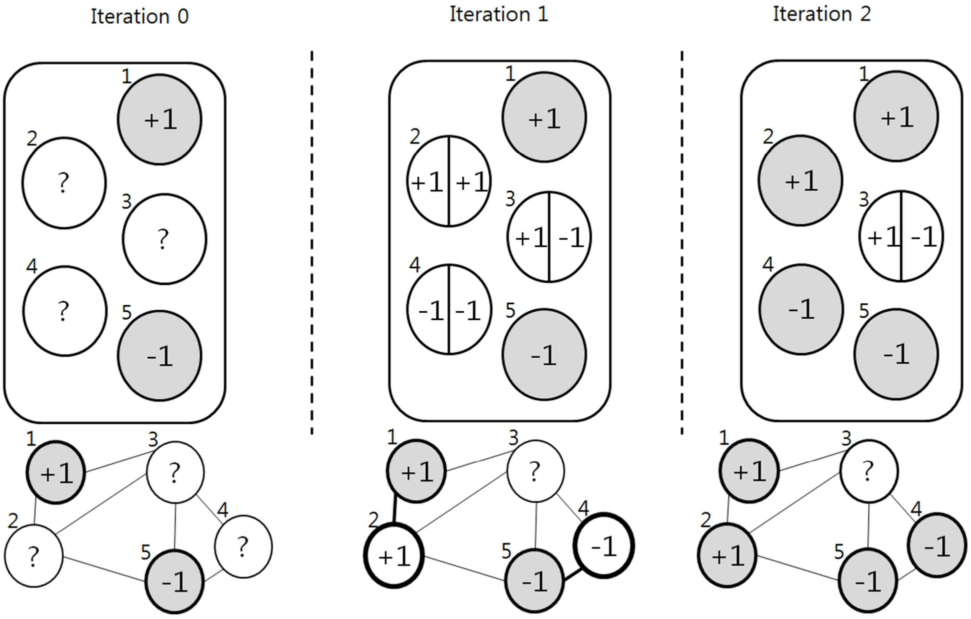


Figure 4. Schematic description for SSL Co-training: In the beginning (iteration 0), the two data points x1 and x5 belong to the labeled set $L = \{(x_1\ 1), (x_2, -1)\}$ and the labels are given $y_1 = +1$ and $y_5 = -1$, respectively. And x_2, x_3 and x_4 belong to the unlabeled dataset $U = \{x_2, x_2, x_3\}$. After training (iteration 1), the predicted labels for those three data points are given by F_1 and F_2. For x_2, the two classifiers agree on labeling $y_2^1 = y_2^2 = +1$, thus its pseudo-label becomes $x_2 = 1$. Likewise, x_4 obtains the pseudo-label $y_4 = -1$. However, the two classifiers disagree on labeling for x_3: $y_3^1 = +1$ but $y_3^2 = -1$. Therefore, x_3 is a boosted sample by the definition of the proposed algorithm, and remains unlabeled. In the next iteration (iteration 2), the labeled dataset is increased with the two pseudo-labeled data points $L = \{(x_1, +1), (x_2, +1), (x_3, -1), (x_4, -1)\}$, and the unlabeled data set is decreased to U=$\{x_3\}$. Similarly with the previous iteration, F_1 and F_2 provide x_3 with the predicted labels $y_3^1 = +1$ and $y_3^2 = -11$, respectively. But again, they fail to agree on labeling for x_3. Since the number of boosted sample is same as the previous iteration, the algorithm stops.

4. Experiments

Data, Performance Measurement, and Experimental Setting

The breast cancer survivability dataset (1973–2003) from SEER is used for experiment, which is an initiative of the National Cancer Institute and is the premier source for cancer statistics in the United States (http://www.seer.cancer.gov) [18]. SEER claims to have one of the most comprehensive collections of cancer statistics. It includes incidence, mortality, prevalence, survival, lifetime risk, and statistics by race/ethnicity. The data consists of 162,500 records with 16 predictor features and 1 target class variable. There are 16 features: tumor size, number of nodes, number of primaries, age at diagnosis, number of positive nodes, marital status, race, behavior code, grade, extension of tumor, node involvement, histologicalTypeICD, primary site, site specific surgery, radiation, and stage. The target variable "survivability" of SEER dataset is a binary categorical feature with values '-1' (not survived) or +1 (survived). Table 1 summarizes the features and their descriptions. The breast cancer survival dataset contains 128,469 positive cases and 34,031 negative cases. To avoid the difficulties in model learning caused by the large-sized and class-imbalanced dataset, 40,000 data points for the training set and 10,000 for the test set are randomly drawn without replacement. The equipoise dataset of 50,000 data points is eventually divided into ten groups, and for each set of which five-fold cross validation is used.

Table 1. Prognostic elements of breast cancer survivability

No.	Features	Description	No.	Features	Description
1	Stage	Defined by size of cancer tumor and its spread	9	Site-Specific Surgery	Information on surgery during first course of therapy, whether cancer-directed or not.
2	Grade	Appearance of tumor and its similarity to more- or less-aggressive tumors	10	Radiation	None, Beam Radiation, Radioisotopes, Refused, Recommended, etc.
3	Lymph Node Involvement	None, (1–3) Minimal, (4–9) Significant, etc.	11	Histological Type	Form and structure of tumor
4	Race	Ethnicity: White, Black, Chinese, etc.	12	Behavior Code	Normal or aggressive tumor behavior is defined using codes.
5	Age at Diagnosis	Actual age of patient in years	13	Number of Positive Nodes Examined	When lymph nodes are involved in cancer, they are called positive.
6	Marital Status	Married, Single, Divorced, Widowed, Separated	14	Number of Nodes Examined	Total nodes (positive/negative) examined
7	Primary Site	Presence of tumor at particular location in body. Topographical classification of cancer.	15	Number of Primaries	Number of primary tumors (1–6)
8	Tumor Size	2–5 cm; at 5 cm prognosis worsens	16	Clinical Extension of Tumor	Defines spread of tumor relative to breast
17	Survivability	Target binary variable defines class of survival of patient.			

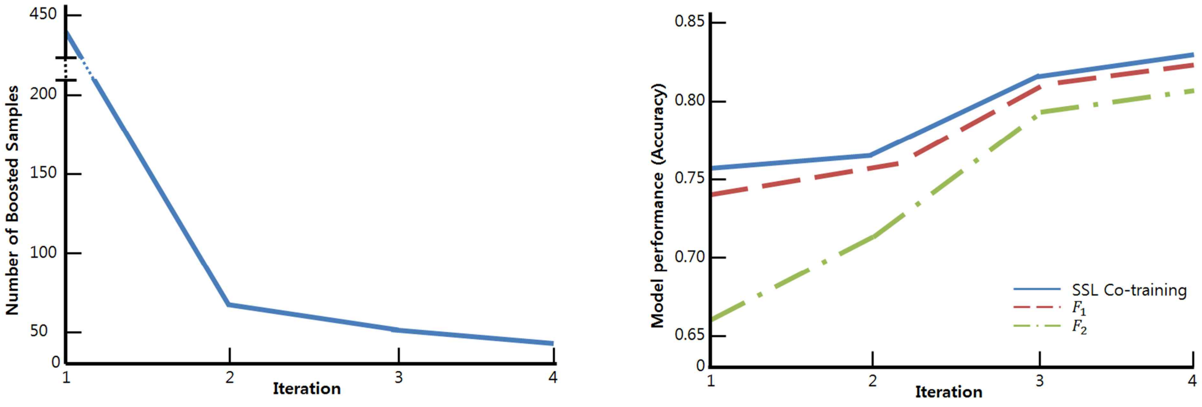
As performance measures, accuracy and the area under the ROC curve (AUC) are used [21][39]. Accuracy is a measure of the total number of correct predictions when the value of classification-threshold is set to 0. On the other hand, AUC assess the overall value of a classifier which is a threshold-independent measure off model performance based on the receiver operating characteristic (ROC) curve which plots the tradeoffs between sensitivity and 1−specificity for all possible values of threshold.

Four representative models, ANN, SVM, SSL and SSL-Co training, are used to perform classification on breast cancer survivability. The model parameters are searched over the following ranges for the respective models. For ANN, the number of 'hidden nodes' and the 'random seed' for initial weights are searched over hidden-node = {3, 6, 9, 12, 15} and random-seed = {1, 3, 5, 7, 10} [24]. For SVM, the values for the RBF kernel width 'Gamma' and the loss penalty term 'C' are selected by searching the ranges of C = {0.2, 0.4, 0.6, 0.8, 1} and Gamma = {0.0001, 0.001, 0.01, 0.1, 1} [27]. For SSL and SSL-Co training models, the values for the number of neighbors 'k' and the tradeoff parameter 'Mu' between smoothness condition and loss condition in (1) are searched over k = {3, 7, 15, 20, 30} and Mu = {0.0001, 0.01, 1, 100, 1000}, respectively.

Results

SSL Co-training for each of the 10 datasets proceeded its iterations between 3 and 5. Figure 5 presents a typical changes in the number of boosted samples and AUC over iterations. The number of boosted samples decreases as iteration proceeds as shown in Figure 5(a) while the AUC performance in Figure 5(b) increases thanks to enhancement of the labeled data set by pseudo-labeled data points. Note that the increasing patterns in AUC of the two member models F_1 and F_2 present the success of co-training between them: F_1helps to lift the performance of F_2 and vice versa.

Table 2 shows the comparison results among ANN, SVM, SSL and SSL Co-training in terms of accuracy and AUC. For each of the four models, the best performance was selected by searching over the respective model-parameter space. For the 10 datasets, the best performance among the four models is marked in boldface. In accuracy, SSL Co-training showed outstanding performance with an average accuracy of 0.76 while SSL was ranked as the second best. In AUC, SSL Co-training produced an average AUC of 0.81 leading the three models although a comparable performance is achieved by SVM as well. Figure 6 summarizes the performances of the four models on two radar graphs.

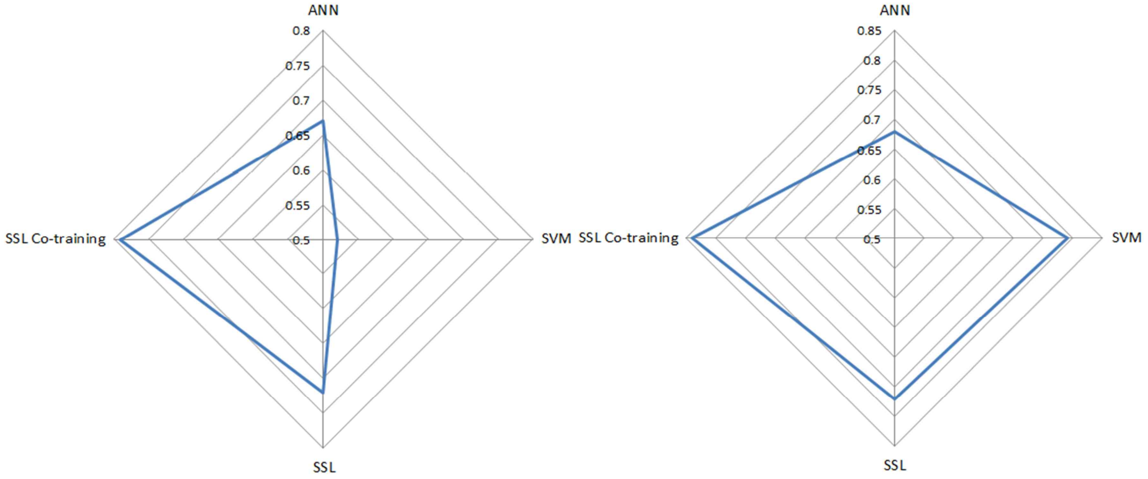


(a) **(b)**

Figure 5. The changes during iterations of SSL Co-training: (a) the number of boosted samples and (b) AUC

Table 2. Performance Comparison among ANN, SVM, SSL and SSL Co-training for the 10 datasets

Data Set	Accuracy				AUC			
	ANN	**SVM**	**SSL**	**SSL Co-training**	**ANN**	**SVM**	**SSL**	**SSL Co-training**
1	0.66	0.52	0.72	**0.77**	0.68	0.79	0.77	**0.84**
2	0.67	0.52	0.72	**0.79**	0.72	0.79	0.79	**0.82**
3	0.62	0.50	0.70	**0.76**	0.68	**0.80**	0.78	0.78
4	0.67	0.51	0.68	**0.75**	0.72	0.79	0.76	**0.81**
5	0.64	0.52	0.71	**0.77**	0.66	**0.82**	0.78	**0.82**
6	0.62	0.52	0.71	**0.76**	0.68	0.78	0.77	**0.83**
7	0.63	0.51	0.69	**0.77**	0.67	0.79	0.77	**0.83**
8	0.69	0.51	0.73	**0.76**	0.73	**0.82**	0.80	**0.82**
9	0.66	0.52	0.70	**0.74**	0.71	**0.81**	0.78	0.78
10	0.64	0.51	0.73	**0.77**	0.73	**0.81**	0.80	**0.81**
Avg.	**0.65**	**0.51**	**0.70**	**0.76**	**0.70**	**0.80**	**0.78**	**0.81**



(a) **(b)**

Figure 6. Performance Comparison among ANN, SVM, SSL and SSL Co-training: (a)Accuracy and (b) AUC

5. Conclusion

In prediction of cancer survivability, obtaining more patient data with labels of either "survived" or "not survived" becomes an important issue since better prediction of a predictive model can be achieved based on them. In practice, however, there are many obstacles when collecting the patient labels because of limitations of time, cost, and conflicts in confidentiality. Therefore, researchers' interests have been attracted to a predictive model that can utilize unlabeled patient data as well which are relatively abundant. In the light of that, SSL has been highlighted as a promising candidate. However, the fact that "the more labeled data, the better prediction" is yet applied to SSL since it is a learning algorithm guided by information contained in the labeled data like other machine learning algorithms. To compensate the lack of labeled data, therefore, SSL Co-training was proposed in this paper. The proposed algorithm generates pseudo-labels as a result of co-training among multiple SSL member models, assigns them to unlabeled data, and eventually uses them as if they are labeled. As the process iterates, labeled data increase and thus the prediction performance of SSL increases. Empirical validation of SSL Co-training on SEER breast cancer database showed successful performance compared with the most representative machine learning algorithms such as ANN, SVM, and ordinary SSL. Using pseudo-labeled patient data together with labeled and unlabeled ones will improve the technical quality of prognosis study on cancer survivability, and the resulting influence is expected to be an aid to provide a better treatment for cancer patients.

The proposed SSL Co-training is still in early development step. Therefore, a series of further study should be accompanied in the near future. First, related to the composition of the member models to co-train, the immediate issues of how to determine the member size and how to make them diverse will be further researched. Second, related to pseudo-labeling process, more sophisticated methods on how to set the cutoff value and how to provide confidence on labeling will be studied.

Acknowledgments

This work would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from National Research Foundation of the Korean Government (2012-0000994)

References

[1] Cancer Facts & Figures 2010, American Cancer Society. Atlanta, 2010.

[2] NC Institute. Breast Cancer Statistics, USA, 2010, National Cancer Institute, 2010, http://www.cancer.gov/cancertopics/types/breast (Accessed: 11 July 2011).

[3] Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, Vol. 23, 2007, pp. 30-37.

[4] Khan U, Shin H, Choi JP, Kim M. wFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability, In: Roddick J F, Li J, Christen P, Kennedy PJ, editors. The Proceedings of the Seventh Australasian Data Mining Conference Glenelg, South Australia, 2008, pp. 141-152.

[5] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, Vol. 34, 2005, pp. 113-127.

[6] Brenner H, Gefeller O, Hakulinen T. A computer program for period analysis of cancer patient survival. European Journal Cancer Vol. 38, 2002, pp. 690-695.

[7] Amir E, Evans DGR, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. Journal of Medical Genetics, Vol. 40, 2003, pp. 807 - 814.

[8] Zhong S. Semi-supervised model-based document clustering: A comparative study. Machine Learning, Vol. 65, 2006, pp. 3-29.

[9] Zhu X. Semi-Supervised Learning with Graphs, Phd. Thesis, School of Computer Science, Carnegie Mellon University, May 2005.

[10] Subramanya A, Bilmes J. Soft-Supervised Learning for Text Classification, In: The Proceedings of the Conference on Empirical Methods in Natural Language Processing Honolulu, Hawaii, 2008, pp. 1090-1099.

[11] Andoy RK, Zhangz T. A High-Performance Semi-Supervised Learning Method for Text Chunking, In: Knight K, Ng HT, Oflazer K, editors. The Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Ann Arbor, Michigan, 2005, pp. 1-9.

[12] Wei L, Keogh E. Semi-Supervised Time Series Classification, In: The Proceedings of the 12th international conference on Knowledge discovery and data mining Philadelphia(KDD 2006), USA, 2006, pp. 748–753.

[13] Bair E, Tibshirani R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data, PLoS Biology, Vol. 2, 2004, pp. 0511-0522.

[14] Gong YC, Chen CL. Semi-supervised Method for Gene Expression Data Classification with Gaussian Fields and Harmonic Functions, In: The Proceedings of 19th International Conference on Pattern Recognition Tampa, FL, 2008, pp. 1-4.

[15] Morsillo N, Pal C, Nelson R. Semi-Supervised Learning of Visual Classifiers from Web Images and Text, In: Boutilier C, editor. The Proceedings of the 21st international joint conference on Artificial intelligence Pasadena, California, USA, 2009, pp. 1169-1174.

[16] Celikyilmaz A, Thint M, Huang Z. A Graph-based Semi-Supervised Learning for Question-Answering, In: The Proceedings of the 47th Annual Meeting of Annual Meeting of the Association for Computational Linguistics Singapore, 2009, pp. 719–727.

[17] Liu R, Zhou J, Liu M. Graph-based Semi-supervised Learning Algorithm for Page Classification, In: The Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications China, IEEE Computer Society, 2006, pp. 856 – 860.

[18]SEER, Surveillance, Epidemiology and End Results program National Cancer Institute, 2010, http://www.seer.cancer.gov (Accessed: 11 July 2011).

[19]N.B.C. Foundation, What is Breast Cancer?, National Breast Cancer Foundation, Inc, 2010, http://www.nationalbreastcancer.org/?aspxerrorpath=/about-breast-cancer/what-is-breast%20cancer.aspx (Accessed: 11 July 2011).

[20] Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics Vol. 2, 2006, pp. 59-78.

[21] Thongkam J, Xu G, Zhang Y, Huang F. Breast Cancer Survivability via AdaBoost Algorithms, In: Warren JR, Yu P, Yearwood J, Patrick JD, editors. The Proceedings of the second Australasian workshop on Health data and knowledge management Wollongong, NSW, Australia, 2008, pp. 55-64.

[22] Thongkam J, Xu G, Zhang Y, Huang F. Towards breast cancer survivability prediction models through improving training space. Expert Systems with Applications, Vol. 36, 2009, pp. 12200–12209.

[23] Peterson C, Söderberg B. Modern heuristic techniques for combinatorial problems, In: Sons JW. (Ed.), Artificial Neural Networks New York, USA, 1993, pp. 197-242.

[24] Abraham A. Artificial Neural Networks, In: Sydenham P, Thorn R, editors. Handbook for Measurement Systems Design, London, 2005.

[25] Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, Vol. 26, 2006, pp. 159–190.

[26] Cardoso JS, Cardoso MJ. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artificial Intelligence in Medicine Vol. 40, 2007, pp. 115-126.

[27] Shin H, Cho S. Neighbourhood Property-Based pattern selection for support vector machines. Neural Computation Vol. 19, 2007 pp. 816-855.

[28] Schölkopf B, Smola AJ. Learning with Kernels, The MIT Press, Cambridge, England, 2002.

[29] Choi I, Shin H. Semi-supervised Learning with Ensemble Learning and Graph Sharpening, In: Colin F, Kim DS, Lee SY, editors. The Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning, Daejeon, South Korea, 2008, pp. 172-179.

[30] He J, Carbonell J, Liu Y. Graph-Based Semi-Supervised Learning as a Generative Model, In: Veloso MM. editor. The Proceedings of the 20th international joint conference on Artificial intelligence, Hyderabad, India, 2007, pp. 2492-2497.

[31] Chapelle O, Schölkopf B, Zien A. Semi-Supervised Learning, The MIT Press, Cambridge, England, 2006, pp. 3-14.

[32] Wang J. Efficient large margin semi-supervised learning. Journal of Machine Learning Research Vol. 10, 2007, pp. 719-742.

[33] Zhu X. Semi-Supervised Learning Literature Survey, Computer Sciences TR 1530 Madison, University of Wisconsin, 2008.

[34] Shin H, Hill NJ, Lisewski AM, Park JS. Graph sharpening. Expert Systems with Applications, Vol. 37, 2010, pp. 7870-7879.

[35] Belkin M, Matveeva I, Niyogi P. Regularization and Semi-supervised Learning on Large Graphs, In: Lecture Notes in Computer Science, Springer, Vol. 3120, 2004, pp. 624-638.

[36] Chapelle O, Weston J, Schölkopf B. Cluster Kernels for Semi-Supervised Learning, In: Advances in Neural Information Processing Systems, The MIT Press, Cambridge, England, 2003, pp. 585-592.

[37] Shin H, Lisewski AM, Lichtarge O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics, Vol. 23, 2007, pp. 3217-3224.

[38] Altman D, Bland M. Diagnostic tests; 1: Sensitivity and specificity. British Medical Journal (BMJ), Vol. 308, 1994, pp. 1552–1552.

[39] Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology, Vol. 43, 2006, pp. 1223–1232.

[40] Sheldon MR, Fillyaw MJ, Thompson WD. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs, Physiotherapy Research International, Vol.1, No.4, 1996, pp. 221-228

[41] Shin H, Kim D, Park K and Ali A. Breast Cancer Survivability Prediction with Surveillance, Epidemiology, and End Results Database, TBC, Seoul, Korea, 2011