On Selecting Helpful Unlabeled Data for Improving Semi-Supervised Support Vector Machines*

Le Thanh-Binh and Kim Sang-Woon
Department of Computer Engineering, Myongji University, Yongin, 449-728 South Korea

Keywords: Semi-Supervised Learning, Support Vector Machines, Semi-Supervised Support Vector Machines.

Abstract: Recent studies have demonstrated that Semi-Supervised Learning (SSL) approaches that use both labeled and unlabeled data are more effective and robust than those that use only labeled data. However, it is also well known that using unlabeled data is not always helpful in SSL algorithms. Thus, in order to select a small amount of helpful unlabeled samples, various selection criteria have been proposed in the literature. One criterion is based on the prediction by an ensemble classifier and the similarity between pairwise training samples. However, because the criterion is only concerned with the distance information among the samples, sometimes it does not work appropriately, particularly when the unlabeled samples are near the boundary. In order to address this concern, a method of training semi-supervised support vector machines (S3VMs) using selection criterion is investigated; this method is a modified version of that used in SemiBoost. In addition to the quantities of the original criterion, using the estimated conditional class probability, the confidence values of the unlabeled data are computed first. Then, some unlabeled samples that have higher confidences are selected and, together with the labeled data, used for retraining the ensemble classifier. The experimental results, obtained using artificial and real-life benchmark datasets, demonstrate that the proposed mechanism can compensate for the shortcomings of the traditional S3VMs and, compared with previous approaches, can achieve further improved results in terms of classification accuracy.

1 INTRODUCTION

In semi-supervised learning (SSL) approaches, a large amount of unlabeled data (U), together with labeled data (L), is used to build better classifiers. That is, SSL exploits the samples of U in addition to the labeled counterparts in order to improve the performance of a classification task, which leads to a performance improvement in the supervised learning algorithms with a multitude of unlabeled data. However, it is also well known that using U is not always helpful for SSL algorithms. In particular, it is not guaranteed that adding U to the training data (T), i.e. $T = L \cup U$, leads to a situation in which the classification performance can be improved (Ben-David, S. et al., 2008; Lu, T., 2009; Zhu, X., 2006). Therefore, if more is known about the confidence levels involved in classifying U, informative data could be chosen and included easily when training base classifiers. Furthermore, if a large amount of unlabeled samples could be added to the training set, then the number of training samples could be expanded effectively. Using large and strong training samples may lead to creating a strongly learned classifier.

From this perspective, in order to select a small amount of helpful unlabeled data, various selecting techniques have been proposed in the literature, including the self-training (McClosky, D. et al., 2008; Rosenberg, C. et al., 2005), co-training (Blum, A. and Mitchell, T., 1998; Du, J. et al., 2011), cluster-then-label (Singh, A. et al., 2008; Goldberg, A. B. et al., 2009; Goldberg, A. B., 2010), simply recycled strategy in SemiBoost (Mallapragada, P. K. et al., 2009), incrementally reinforced semi-supervised MarginBoost (SSMB) (Le, T. -B. and Kim, S. -W., 2012), and other criteria used in active learning (AL) algorithms (Dagan, I. and Engelson, S. P., 1995; Riccardi, G. and Hakkani-Tur, D., 2005; Kuo, H. -K. J. and Goel, V., 2005; Leng, Y. et al., 2013). For example, in SemiBoost, Mallapragada *et al.* measured the pairwise similarity in order to guide the selection of a subset of U for each iteration and to assign (pseudo) labels to them. That is, they first

*This work was supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2012R1A1A2041661).

Le T. and Kim S..
On Selecting Helpful Unlabeled Data for Improving Semi-Supervised Support Vector Machines.
DOI: 10.5220/0004810500480059
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 48-59
ISBN: 978-989-758-018-5

computed the confidence of all U samples based on the prediction made by an ensemble classifier and the similarity among the samples of $L \cup U$. Then, they selected a few samples with higher confidence to retrain the ensemble classifier together with L. The selecting-and-training step was repeated for the number of iterations or until a termination criterion was met.

On the other hand, support vector machines (SVMs) (Vapnik, V., 1995) are considered to be strong and successful classifiers in pattern recognition (PR). Unlike traditional classification models, such as Bayesian decision rules, SVMs minimize the upper bound of the generalization error by maximizing the margin between the separating hyperplane and training data. Hence, SVMs are a distribution-free model that can overcome the problems of poor statistical estimation and small sample sizes. SVMs also achieve greater empirical accuracy and better generalization capabilities than other standard supervised classifiers. With regard to combining SVMs with SSL strategies, numerous models use unlabeled samples to improve the classification performance, including semi-supervised support vector machines (S3VMs) (Bennett, K. P. and Demiriz, A., 1998), transductive support vector machines (TSVMs) (Joachims, T., 1999b), EM algorithms with generative mixture models (Nigam, K. et al., 2000), Bayesian S3VMs (Chakraborty, S., 2011), help-training (which is a variant of the self-training) S3VMs (Adankon, M. M. and Cheriet, M., 2011), hybrid S3VMs (Jiang, Z. et al., 2013), and S3VM-us (semi-supervised support vector machines with unlabeled instances selection) (Li, Y. -F. and Zhou, Z. -H., 2011).

Among these combined approaches, the semi-supervised support vector machines (S3VMs) (Bennett, K. P. and Demiriz, A., 1998; Chapelle, O. et al., 2006) and the transductive support vector machines (TSVMs) (Joachims, T., 1999b) are the most popular approaches for utilizing unlabeled data. In particular, S3VMs are constructed using a mixture of L (training set) and U (working set) data, where the objective is to assign class labels to the working set. Therefore, when the working set is empty, the S3VM becomes the standard SVM model. In contrast, when the training set is empty, it becomes an unsupervised learning approach (Bennett, K. P. and Demiriz, A., 1998; Joachims, T., 1999b). Consequently, when both the training and working sets are not empty, SSL strategies can be used. In this case, the information from U can be helpful for the training process. Moreover, without labels, the cost of extracting U samples may be lower than that of providing more L samples. Therefore, S3VMs create a richness of opportunity for many PR researchers.

The combination of helpful U samples with L data increases the likelihood of more accurate classification; however, the determination of estimated labels for U often leads to a fault. If this fails, the added U samples with incorrect labels not only decrease the accuracy of the classification but also increase the difficulty in choosing a decision function. From this perspective, in order to complement the weakness of S3VM, various techniques, such as SemiBoost (Mallapragada, P. K. et al., 2009), conjugate function strategy (Sun, S. and Shawe-Taylor, J., 2010), S3VM-us (Li, Y. -F. and Zhou, Z. -H., 2011), incrementally reinforced selection strategy (Le, T. -B. and Kim, S. -W., 2012), manifold-preserving graph reduction (Sun, S. et al., 2014), etc., have been proposed in the literature. In SemiBoost, for example, the confidence value of $x_i \in U$ is computed using two quantities, i.e. p_i and q_i, which are measured using the pairwise similarity between x_i and other U and L samples. However, when x_i is near the boundary between two classes, the value is computed using U only, without referring to L. Consequently, the value might be inappropriate for selecting helpful samples. In order to address problem, a modified technique that minimizes the errors in estimating the labels of U is investigated.

This modification is motivated using the observation that, for samples $x_i \in U$ that are near the boundary between the positive class of L (L^+) and the negative class of L (L^-), three terms that comprise the selection criterion of SemiBoost are reduced to one term, which only depends on U. That is, two of the three terms, which are measured using L^+ and L^-, respectively, are changed to zero or nearly zero. From this observation, the balance between the impacts of the labeled and pseudo-labeled data is used when computing the confidence values. The difference between both criteria is two-fold: the first difference is that, for the original criterion of SemiBoost, the confidence values are computed using the quantities of p_i and q_i only, whereas for the modified criterion, they are measured using estimates of the conditional class probabilities as well as the quantities of p_i and q_i. The second difference is the method of labeling the selected samples: in the original scheme, the label of $x_i \in U$ is predicted using a $sign(p_i - q_i)$, while in the modified scheme, this is predicted by referring to the probability estimates as well as p_i and q_i.

The main contribution of this paper is the demonstration that the classification accuracy of S3VM can be improved using a modified criterion when selecting unlabeled samples and predicting their labels. Furthemore, a comparison of the classification performance between the proposed S3VM and the traditional ones was performed empirically. In particu-

lar, some critical questions concerning the strategies employed in the present work were investigated, including *what are the features of the original S3VM and SemiBoost that lead to the lower classification accuracy?* and *why is the proposed modified criterion better than the original?*

The remainder of the paper is organized as follows. In Section 2, after providing a brief introduction to S3VMs, an explanation for the use of selection criterion in the SemiBoost algorithm is provided. Then, in Section 3, a method of improving S3VMs through utilizing the modified criterion for selecting a small amount of helpful unlabeled samples is presented. In Sections 4 and 5, the experimental setup and results obtained using the experimental benchmark data are presented, respectively. Finally, in Section 6, the concluding remarks and limitations that deserve further study are presented.

2 RELATED WORK

In this section, S3VM and SemiBoost, which are closely related to the present empirical study, are briefly reviewed. The details of the algorithms can be found in the related literature (Vapnik, V., 1995; Bennett, K. P. and Demiriz, A., 1998; Mallapragada, P. K. et al., 2009).

2.1 S3VM and TSVM

A set of *nl* training pairs ($L = \{(x_1, y_1), \cdots, (x_{nl}, y_{nl})\}$, $x_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}$) and a set of *nu* unlabeled samples ($U = \{x_1, \cdots, x_{nu}\}$ and $x_j \in \mathbb{R}^d$) are considered. Referring to (Vapnik, V., 1995), SVMs have a decision function $f_\theta(\cdot)$, which is defined as $f_\theta(x) = w \cdot \Phi(x) + b$, where $\theta = (w, b)$ denotes the parameters of the classifier model, $w \in \mathbb{R}^d$ is a vector that determines the orientation of the discriminating hyperplane, and $b \in \mathbb{R}$ is a bias constant such that $b/\|w\|$ represents the distance between the hyperplane and origin. Also, $\Phi : \mathbb{R}^d \to F$ is a nonlinear feature mapping function, which is often implemented implicitly using the kernel trick.

When denoting η_i as the loss for x_i, the quadratic programming formulation is defined as follows:

$$\begin{aligned} &min \ \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\eta_i \\ &s.t. \ \ y_i f_\theta(x_i) + \eta_i \geq 1, \eta_i \geq 0, i = 1, \cdots, nl, \end{aligned} \tag{1}$$

where $C > 0$ is a fixed penalty regularization parameter, which is determined via trial and error (Vapnik, V. and Chervonenkis, A. I., 1974), (Vapnik, V., 1982), (Vapnik, V., 1995). In particular, S3VM is defined as follows (Bennett, K. P. and Demiriz, A., 1998):

$$\begin{aligned} &min \ \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\eta_i + C^*\sum_{j=1}^{nu}\eta_j \\ &s.t. \ \ y_i f_\theta(x_i) + \eta_i \geq 1, i = 1, \cdots, nl, \\ &\quad\ |f_\theta(x_j)| \geq 1 - \eta_j, j = 1, \cdots, nu. \end{aligned} \tag{2}$$

S3VMs are an expansion of SVMs using an SSL strategy, while TSVMs use the transductive learning approach. Given a set of *nl* training pairs (L) and a (unlabeled) set of *nt* test samples in test set (T_U), the goal is to determine the pairs that an SVM trained on $L \cup (T_U \times Y^*)$ can use to yield the largest margin from the possible binary estimated label vectors $Y^* = (y_{nl+1}, \cdots, y_{nl+nt})$. This is a combinatorial problem, but it can be approximated (see (Vapnik, V., 1995)) to locating an SVM that separates the training set under constraints, which forces the test unlabeled samples to be as far as possible from the margin. This can be written as follows:

$$\begin{aligned} &min \ \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\eta_i + C^*\sum_{j=1}^{nt}\eta_j \\ &s.t. \ \ y_i f_\theta(x_i) + \eta_i \geq 1, \eta_i \geq 0, i = 1, \cdots, nl, \\ &\quad\ |f_\theta(x_j)| \geq 1 - \eta_j, j = 1, \cdots, nt. \end{aligned} \tag{3}$$

This minimization problem is equivalent to minimizing $\mathcal{L}$, which is defined as follows:

$$\begin{aligned} \mathcal{L} \ \equiv \ \frac{1}{2}\|w\|^2 \ &+ \ C\sum_{i=1}^{nl}\mathcal{H}_1(y_i f_\theta(x_i)) \\ &+ \ C^*\sum_{j=1}^{nt}\mathcal{H}_1(|f_\theta(x_j)|), \end{aligned} \tag{4}$$

where $\mathcal{H}_1(\cdot)$ is the Hinge loss function defined as follows:

$$\mathcal{H}_1(\gamma) = \begin{cases} 1-\gamma, & if \ \ \gamma < 1 \\ 0 & otherwise. \end{cases} \tag{5}$$

For $C^* = 0$ in (4), the standard SVM optimization problem is obtained. For $C^* > 0$, the U data that are inside the margin are penalized. This is equivalent to using the Hinge loss on U as well, but it is assumed that the label of the unlabeled example x_i is $y_i = sign(f_\theta(x_i))$. In order to solve (4), Joachims (Joachims, T., 1999b) proposed an efficient local search algorithm that is the basis of SVM^{Light} (Joachims, T., 1999a).

2.2 SemiBoost

The goal of SemiBoost (Mallapragada, P. K. et al., 2009), which is a boosting framework for SSL, is to iteratively improve the performance of a supervised

learning algorithm ($\mathcal{A}$) by regarding it as a black box, using U and pairwise similarity. In order to follow the boosting idea, SemiBoost optimizes performance through minimizing the objective loss function defined as follows (see Proposition 2 (Mallapragada, P. K. et al., 2009)):

$$\overline{F_1} \leq \sum_{i=1}^{nu}(p_i+q_i)(e^{2\alpha}+e^{-2\alpha}-1) - \sum_{i=1}^{nu}2\alpha h_i(p_i-q_i), \quad (6)$$

where $h_i(=h(x_i))$ is the classifier learned by $\mathcal{A}$ at the iteration, α is the weight for combining h_i's, and

$$p_i = \sum_{j=1}^{nl} S_{i,j}^{ul} e^{-2H_i}\delta(y_j,1) + \frac{K}{2}\sum_{j=1}^{nu} S_{i,j}^{uu} e^{H_j-H_i},$$
$$q_i = \sum_{j=1}^{nl} S_{i,j}^{ul} e^{2H_i}\delta(y_j,-1) + \frac{K}{2}\sum_{j=1}^{nu} S_{i,j}^{uu} e^{H_i-H_j}. \quad (7)$$

Here, $H_i(=H(x_i))$ denotes the final combined classifier and S denotes the pairwise similarity. For all x_i and x_j of the training set, for example, S can be computed using as follows:

$$S(i,j) = exp(-\|x_i - x_j\|_2^2/\sigma^2), \quad (8)$$

where σ is the scale parameter controlling the spread of the function. In addition, S^{lu} (and S^{uu}) denotes the $nl \times nu$ (and $nu \times nu$) submatrix of S. Also, S^{ul} and S^{ll} can be defined correspondingly; the constant K, which is computed using $K = |L|/|U| = nl/nu$, is introduced to weight the importance between L and U; and $\delta(a,b) = 1$ when $a = b$ and 0 otherwise.

The quantities of p_i and q_i can be interpreted as the confidence in classifying $x_i \in U$ into a positive class ($\{+1\}$) and negative class ($\{-1\}$), respectively. Using these settings, p_i and q_i can be used to guide the selection of U samples at each iteration using the confidence measurement $|p_i - q_i|$, as well as to assign the pseudo class label $sign(p_i - q_i)$. The procedure of selecting strong samples from U using confidence levels, which is referred to as a *sampling* function, is summarized as follows.

From (7), the difference in values between p_i and q_i can be formulated as follows:

$$p_i - q_i = \sum_{j=1}^{nl} S_{i,j}^{ul} e^{-2H_i}\delta(y_j,1) - \sum_{j=1}^{nl} S_{i,j}^{ul} e^{2H_i}\delta(y_j,-1) + \frac{C}{2}\sum_{j=1}^{nu} S_{i,j}^{uu}(e^{H_j-H_i} - e^{H_i-H_j}). \quad (9)$$

Algorithm 1: Sampling.

Input: Labeled data (L) and unlabeled data (U).
Output: Selected unlabeled data (U_s).
Procedure: Repeat the following steps to select U_s from U.

1. For each sample of U, compute classification confidence levels ($\{|p_i - q_i|\}_{i=1}^{nu}$) using (7).
2. After sorting the levels $|p_i - q_i|$ in descending order, choose a small portion from the top of the unlabeled data (e.g. 10% top) as U_s, according to the confidence levels.
3. Update the estimated label for any selected sample x_i by $sign(p_i - q_i)$.

End Algorithm

By substituting $L^+ \equiv \{(x_i,y_i)|y_i = +1, i = 1,\cdots,nl^+\}$ and $L^- \equiv \{(x_i,y_i)|y_i = -1, i = 1,\cdots,nl^-\}$ as the L samples in class $\{+1\}$ and class $\{-1\}$, respectively, (9) can be represented as follows:

$$p_i - q_i = \left(e^{-2H_i}\sum_{x_j\in L^+} S_{i,j}^{ul}\right) - \left(e^{2H_i}\sum_{x_j\in L^-} S_{i,j}^{ul}\right) + \left(\frac{C}{2}\sum_{x_j\in U} S_{i,j}^{uu}(e^{H_j-H_i} - e^{H_i-H_j})\right). \quad (10)$$

Again, by substituting $X_i^+ \equiv e^{-2H_i}\sum_{x_j\in L^+} S_{i,j}^{ul}$ and $X_i^- \equiv e^{2H_i}\sum_{x_j\in L^-} S_{i,j}^{ul}$ in the first two corresponding summations of the similarity distances from $x_i \in U$ to each $x_j \in L$ in class $\{+1\}$ and class $\{-1\}$, the difference in the values between X_i^+ and X_i^- can be considered as the relative measurement for estimating the possibility that x_i belongs to $\{+1\}$ or $\{-1\}$ as follows:

$$X_i^+ - X_i^- < 0 \Rightarrow P(x_i \in \{+1\}) < P(x_i \in \{-1\}),$$
$$X_i^+ - X_i^- > 0 \Rightarrow P(x_i \in \{+1\}) > P(x_i \in \{-1\}). \quad (11)$$

From this representation, it can be seen that if the difference of X_i^+ and X_i^- is nearly zero, then the sample x_i could remain on the boundary of the classifier. Therefore, the classification of x_i is a complicated problem. In order to address this problem, SemiBoost uses the third term in (10), which denotes the relative information (i.e. similarity) between $x_i \in U$ and $x_j \in U$. This may provide more meaningful information for enlarging the margin.

However, providing more data is not always beneficial. If the value obtained using the third term in

(10) is very large or X_i^+ is nearly equal to X_i^-, (10) will generate some erroneous data. In that case, the meaning achieved using the confidence of $X_i^+ - X_i^-$ may be lost and the estimation for x_i will depend on the U data. That is, the L samples do not affect the estimation of x_i label; therefore, the estimated label is unsafe and untrustworthy.

3 PROPOSED METHOD

In this section, in order to overcome the above mentioned weakness, the selection/prediction criterion based on p_i and q_i is modified and, using the modified criterion, a learning algorithm for S3VMs is proposed.

3.1 Quadratic Optimization Problem

First, the focus is on optimizing (2) in order to minimize the quadratic problem to improve the results of S3VMs. Minimizing (2) leads to the generation of an optimized classifier. Let the U_s be a subset of ns samples selected from U that have a high possibility of trust. That is, U is partitioned into two subsets, i.e. the selected U and remaining U ($U = U_s \cup U_r$), where the cardinalities of U_s and U_r are ns and nr, respectively. Thus, the minimum (2) would be divided into two terms represented using brackets as follows:

$$\begin{aligned} min\ \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\eta_i + \left[C^*\sum_{j=1}^{ns}\eta_j + C^*\sum_{k=1}^{nr}\eta_k\right] \\ s.t.\ \ & y_i f_\theta(x_i) + \eta_i \geq 1, \eta_i \geq 0, i = 1, \cdots, nl, \\ & |f_\theta(x_j)| \geq 1 - \eta_j, j = 1, \cdots, nu. \end{aligned} \tag{12}$$

Using the Hinge loss in (5) for TSVMs, minimizing (12) is similar to minimizing $\mathcal{L}$, which is computed as follows:

$$\begin{aligned} \mathcal{L} = \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\mathcal{H}_1(y_i f_\theta(x_i)) \\ & + \left[C^*\sum_{j=1}^{ns}\mathcal{H}_1(|f_\theta(x_j)|) + C^*\sum_{k=1}^{nr}\mathcal{H}_1(|f_\theta(x_k)|)\right]. \end{aligned} \tag{13}$$

From (13), it is easy to observe that a smaller value can be achieved when reinforcing the training set with U_s and its predicted labels. Furthermore, by omitting the term related to the U_r subset from (13), the problem of minimizing $\mathcal{L}$ can be simplified to the minimization of $\mathcal{L}_1$, which is defined as follows:

$$\begin{aligned} \mathcal{L}_1 \equiv \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\mathcal{H}_1(y_i f_\theta(x_i)) \\ & + \left[C^*\sum_{j=1}^{ns}\mathcal{H}_1(|f_\theta(x_j)|)\right]. \end{aligned} \tag{14}$$

Thus, it can be seen that $\mathcal{L}_1 \leq \mathcal{L}$ without losing generality. From this observation, rather than optimizing $\mathcal{L}$, $\mathcal{L}_1$ can be considered as a new quadratic optimization problem. Furthermore, it should be noted that the quadratic problem could be more efficiently optimized through the minimization of each term in (14), not through a summation. Therefore, a modified version of the selection criterion in (10) could be considered. In subsequent sections, the method of adjusting the selection (sampling) function and using it are discussed.

3.2 Modified Criterion

As mentioned previously, using p_i and q_i can lead to incorrect decisions in the selection and labeling steps; this is particularly common when the summation of the similarity measurement from $x_i \in U$ to $x_j \in L$ is too weak, as follows:

$$X_i^+ - X_i^- \ll X_i^u, \tag{15}$$

where $X_i^u \equiv \left(\frac{C}{2}\sum_{x_j \in U} S_{i,j}^{uu}(e^{H_j - H_i} - e^{H_i - H_j})\right)$, or

$$X_i^+ \approx X_i^-. \tag{16}$$

In this situation, the confident measurement is formulated as follows:

$$|p_i - q_i| \simeq |X_i^u|. \tag{17}$$

From (17), it can be observed that the confident measurement of $x_i \in U$ is computed using the distance between x_i and $x_j \in U$, while excluding L. As a consequence, the measurement is determined using U only and, therefore, sometimes it does not function as a criterion for selecting strong samples. In order to avoid this, the criterion of (10) can be improved through balancing the three terms in (10), i.e. X_i^+, X_i^-, and X_i^u. This improvement can be achieved through balancing the three terms through a reduction in the impact of the third term, especially when $X_i^+ \approx X_i^-$. More specifically, in order to reduce the impact, the conditional class probability is estimated with each $x_i \in U$ in this paper. This idea is motivated from the rule of mapping the selected unlabeled sample (x_i) to a predicted label (y_i) being viewed as a procedure for obtaining the estimates of a set of conditional probabilities.

In order to obtain the estimates of the probabilities, a method cited from the LIBSVM library (Chang, C. -C. and Lin, C. -J., 2011) can be considered. Using the probability estimates as a penalty cost, the criterion of (10), i.e. $|p_i - q_i|$, can be modified as follows:

$$|CL(x_i)| = \left|X_i^+ - X_i^- + X_i^u - (1 - P_E(x_i))\right|, \quad (18)$$

where $P_E(x_i)$ denotes the probability estimates and $1 - P_E(x_i)$ corresponds to the percentage of mistakes when labeling x_i. Using (18) as the criterion of selecting strong unlabeled samples, the sampling function described in Section 2.2 can be modified as follows.

Algorithm 2: Modified Sampling.

Input: Labeled data (L) and unlabeled data (U).
Output: Selected unlabeled data (U_s).
Procedure: Repeat the following steps to select U_s from U.

1. For each sample of the available unlabeled data, compute the classification confidence levels $\{|CL(x_i)|\}_{i=1}^{nu}$ using (18).
2. After sorting the levels in descending order, choose a small portion of the top of the unlabeled data (e.g. 10% top) as U_s, according to their confidence levels.
3. Update the estimated label for any selected sample x_i using $sign(CL(x_i))$.

End Algorithm

3.3 Proposed Algorithm

In this section, an algorithm that upgrades the conventional S3VM through the modified criterion for selecting helpful samples from U is presented. The algorithm begins with predicting the labels of U using an SVM classifier trained with L only. After initializing the related parameters, e.g. the kernel function and its related conditions, the confidence levels of U ($\{|CL(x_i)|\}_{i=1}^{nu}$) are calculated using (18). Then, $\{|CL(x_i)|\}_{i=1}^{nu}$ is sorted in descending order. After selecting the samples ranked with the highest confidence levels, combining them with L creates a training set for an S3VM classifier. In training the S3VM classifier, the minimization problem, which corresponds to (4), can be solved through minimization:

$$\begin{aligned}\mathcal{L}_1 = \; & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{nl}\mathcal{H}_1(y_i f_\theta(x_i)) \\ & + C^*\sum_{j=1}^{ns}\mathcal{H}_1\left(sign(CL(x_i))\, f_\theta(x_j)\right),\end{aligned} \quad (19)$$

where $\mathcal{H}_1$ is the Hinge loss function in (5).

Finally, the selection and training steps are repeated while verifying the training error rates of the classifier. The repeated regression leads to an improved classification process and, in turn, provides better prediction of the labels over iterations. Consequently, the best training set, which is composed of L and U_s samples, constitutes the final classifier for the problem.

Based on this brief explanation, an algorithm for improving the S3VM using the modified criterion is summarized as follows, where the labeled and unlabeled data (L and U), cardinality of U_s, number of iterations (e.g. $t_1 = 100$), and type of kernel function and its related constants (i.e. C and C^*), are given as input parameters. As outputs, the labels of all data and the classifier model are obtained:

Algorithm 3: Proposed Algorithm.

Input: Labeled data (L) and unlabeled data (U).
Output: Final classifier (H_f).
Method:
Initialization: Select $U_s^{(0)}$ from U through an SVM trained with L; set the parameters, e.g. C and C^*, and kernel function (Φ); train the first S3VM (H_f) with $L \cup U_s^{(0)}$ and compute the training error ($\varepsilon(H_f)$), using L only.
Procedure: Repeat the following steps while increasing i from 1 to t_1 in increments of 1.

1. Choose $U_s^{(i)}$ from U using the modified sampling function (i.e., Algorithm 2), where the previously trained S3VM is invoked.
2. Train a new S3VM classifier (h_i) using both L and $U_s^{(i)}$, and obtain the training error ($\varepsilon(h_i)$) with L.
3. If $\varepsilon(h_i) \leq \varepsilon(H_f)$, then keep h_i as the best classifier, i.e. $H_f \leftarrow h_i$ and $\varepsilon(H_f) \leftarrow \varepsilon(h_i)$.

End Algorithm

The time complexities of the two algorithms, the SemiBoost (Mallapragada, P. K. et al., 2009) algorithm and the proposed algorithm, can be analyzed and compared as follows. As in the case of SemiBoost algorithm, almost all the processing CPU-time of the proposed algorithm is also consumed in computing the three steps of **Procedure** in **Algorithm 3**. So, the difference in magnitude between the computational complexities of SemiBoost and the proposed algorithm depends on the computational costs associated with the routines of three steps. More specifically, in both algorithms, the three steps are concerned with: (1) sampling a small amount of the unlabeled samples U using the criteria; (2) learning a

Table 1: Comparison of time complexities of the three steps for the SemiBoost algorithm and the proposed algorithm. Here, $|\cdot|$ denotes the cardinality of a data set.

Steps	SemiBoost algorithm	Proposed algorithm
(1) Sampling	$O(\|U\|+\|U\|log\|U\|)$	$O(\|U\|+\|U\|log\|U\|)$
(2) Training	$O(\|L\|+\|S\|)$	$O(\|L\|+\|S\|)$
(3) Updating weights	$O(\|L\|+\|U\|)$	–
(and the best S3VM)	$O(1)$	$O(1)$

weak-learner (and S3VM in **Algorithm 3**) using the labeled data L and the selected samples S; and (3) updating the ensemble classifier with the appropriately estimated weights for SemiBoost, while keeping the best classifier for the proposed algorithm. From this consideration, the time complexities for the steps can be summarized in Table 1.

From Table 1, in the case of repeating the three steps t times, the time complexities of the two algorithms are, respectively, $O(\alpha_1 t)$ and $O(\alpha_2 t)$, where $\alpha_1 = 2|U|+|U|log|U|+2|L|+|S|+1$ and $\alpha_2 = |U|+|U|log|U|+|L|+|S|+1$, and, consequently, $\alpha_1 > \alpha_2$. From this analysis, it can be seen that the required time for SemiBoost is much more sensitive to the cardinalities of the training sets (L and U) and the selected data set (S) than that for the proposed algorithm.

4 EXPERIMENTAL SETUP

In this section, in order to perform experiments for evaluating the proposed approach, experimental data and methods are described first.

4.1 Experimental Data

The proposed algorithm was evaluated and compared with the traditional algorithms. This was accomplished through performing experiments on

the *Image Classification Practical 2011* database [2] , which was published by Vedaldi and Zisserman (Vedaldi, A. and Zisserman, A., 2011). This database contains five groups of image data: *person*, *horse*, *car*, *aeroplane*, and *motorbike*. Each group contains one class $\{+1\}$ and must be separated from the other images, called the background image class $\{-1\}$. The background images (1019/4000) are a different image set that is not involved in the five groups mentioned above. The qualification of all image sets is verified using the PASCAL VOC'07 database (Everingham, M. et al., 2007). The characteristics for each group are summarized in Table 2.

[2]http://www.robots.ox.ac.uk/$\tilde{}$vgg/share/practical-image-classification.htm

Table 2: Characteristics of the PASCAL VOC'07 database used in the experiment. Here, four letter acronym, namely, Aero, Moto, Pers, Car, Hors, and Back represent the Aeroplane, Motorbike, Person, Car, Horse, and Background groups, respectively.

Datasets	Aero	Moto	Pers	Car	Hors	Back
Object #	112	120	1025	376	139	1019
Feature #	4000	4000	4000	4000	4000	4000

4.2 Experimental Methods

In this experiment, each dataset was divided into three subsets, i.e. a labeled training set, labeled test set, and unlabeled data set, with a ratio of 20%: 20%: 60%. The training and test procedures were repeated *ten* times and the results were averaged. The (Gaussian) radial basis function kernel, i.e. $\Phi(x,x') = exp(-(\|x-x'\|_2^2)/2\sigma^2)$, was used for all algorithms. In the S3VM classifier, the two constants, C^* and C, were set to 0.1 and 100, respectively, for simplicity. The same scale parameter (σ), which was found using cross-validation by training an inductive SVM for the entire data set, was used for all methods. The proposed S3VM (hereafter referred to as S3VM-improved) was compared with three types of traditional SVMs, which were TSVM (Joachims, T., 1999b), S3VM (Chang, C. -C. and Lin, C. -J., 2011), and SemiBoost-SVM (SB-SVM) (Mallapragada, P. K. et al., 2009), by selecting the top 10% from U.

5 EXPERIMENTAL RESULTS

The run-time characteristics of the proposed algorithm are reported in the following subsections. Prior to presenting the classification accuracies, the original criterion and modified criterion are compared.

5.1 Comparison of Two Criteria: Original and Modified

Prior to presenting the classification accuracies, the original criterion and modified criterion were compared. First, the following question was investigated: does the modified selection criterion perform better than the original criterion? To answer this question, an experiment on selecting unlabeled samples from U was conducted using the original criterion in (9) and the modified criterion in (18). The experiment was

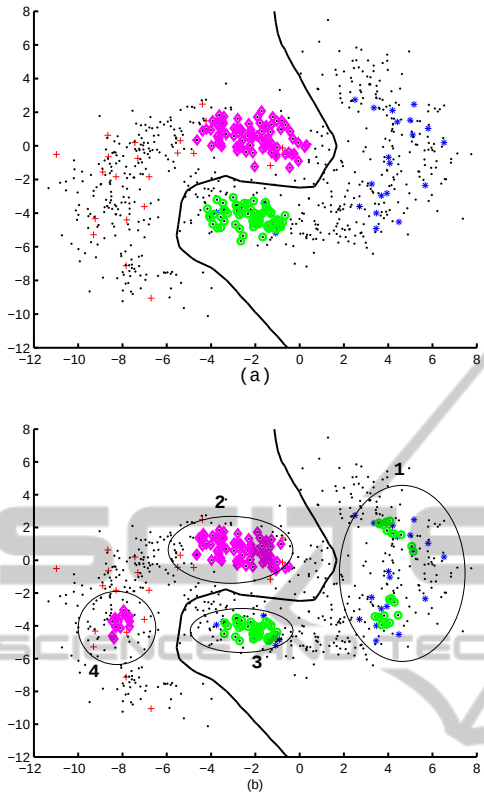

Figure 1: Plots comparing the selected samples with the original criterion (a) and the modified criterion (b) for an artificial dataset. Here, objects in the positive and negative classes are denoted by '$+$' and '$*$' symbols, respectively, in different colors. The selected objects from the two classes are marked with '$\diamond$' and '$\circ$' symbols, respectively from the positive and negative classes, in different colors. The unlabeled data are indicated using a '$\cdot$' symbol.

conducted as follows. First, two confidence values were computed for all U samples with the two criteria in (9) and (18). Second, a subset of U, i.e. U_s (i.e. 10%), was selected referring to the confidence values. Fig. 1 presents a comparison of the two selections achieved using the above experiment for artificial data, which is a two-dimensional, two-class dataset of $[500, 500]$ objects with a banana shaped distribution (Duin,R. P. W. et al., 2004). The data was uniformly distributed along the banana distribution and was superimposed with a normal distribution with a standard deviation $SD = 1$ in all directions. The class priorities are $P(1) = P(2) = 0.5$.

From the figure, it can be observed that the capability of selecting helpful samples for discrimination is generally improved. This is clearly demonstrated in the differences between Fig. 1 (a) and Fig. 1 (b) in the number of selected samples and their geometrical structures. More specifically, for the circled re-

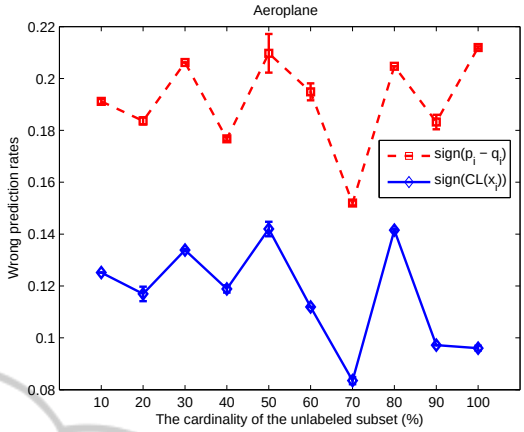

Figure 2: Comparison of the incorrect prediction rates between the original criterion and the modified criterion for the experimental data.

gions #2 and #3, the number of selected points of the modified criterion is smaller than that of the original criterion. In contrast, for the regions #1 and #4, the number of selected points for the modified criterion is larger than that of the original criterion. In the corresponding regions of the latter, there is no selected point. From this observation, it should be noted that the discriminative power of the modified criterion might be better than that of the original criterion.

In order to further investigate this, another experiment was conducted on labeling the unlabeled data using the two selection criteria: a verification of the two predicted labels for each $x_i \in U$ using the two criteria. The experiment was undertaken as follows. First, a subset from U (U_s), for example, the 10% cardinality of U was randomly selected; second, the two labels of all $x_i \in U_s$ predicted using the two techniques in (9) and (18), i.e. $sign(pi - qi)$ and $sign(CL(xi))$, respectively, were compared with their true labels ($y_{U_s} \in \{+1, -1\}$); these two steps were repeated after increasing the cardinality of U_s by 10% until it reached 100%. Fig. 2 presents a comparison of the ten values obtained through repeating the above experiment ten times for the Aeroplane dataset. In the figure, the x-axis denotes the cardinality of U_s and the y-axis indicates the incorrect prediction rates obtained using the two criteria.

From the figure, it can be observed that the prediction capabilities of the original criterion and the modified criterion generally differ from each other; the capability of the modified criterion appears better than that of the original criterion. This is clearly demonstrated in the incorrect prediction rates of the two criteria as represented by the dashed red line with a $\square$ marker and the blue solid line with a $\diamond$ marker for the original and modified criteria, respectively. For all

the datasets and for each repetition, the lower rate was always obtained with the modified criterion described in (18), rather than the original criterion described in (9). That is, in the comparison, the modified criterion always obtained better performance (i.e. the red line with the □ marker is higher than the blue line with the ◇ marker). The same characteristics can be observed in the results from the other datasets. The results of the other datasets are omitted here in order to avoid repetition.

5.2 Comparison of Classification Error Rates between Two Selection Strategies

The following subsection investigates the classification accuracy of the proposed algorithm, i.e. S3VM-improved, using the modified criterion: *is it better (or more robust) than those of the traditional algorithms when the number of selected samples is varied?* In order to answer this question and to assess the accuracy of the two selection strategies in particular, the classification error rates of an SVM classifier implemented with a polynomial kernel function of degree 1 and a regularization parameter ($C = 1$), but designed with different training sets (L and different U_s subsets) were tested and evaluated. Here, the two trained SVMs are the SemiBoost-SVM (SB-SVM) and the proposed improved algorithm (S3VM-improved). That is, the S3VM-improved uses the modified criterion to select helpful samples, while the SB-SVM uses the original criterion used in SemiBoost. The comparison was achieved by gradually increasing the cardinality of U_s from 0% to 100%. A cardinality of 0% indicates that the SVM training used only L, while that of 100% indicates that the SVM training used the entire set of U in addition to L. Fig. 3 presents the comparison of the classification error rates of the two approaches for the Aeroplane dataset. In the figure, the x-axis denotes the cardinality of U_s to be added to L, while the y-axis indicates the error rates obtained with the two S3VMs.

In Fig. 3, the blue solid line with a ◇ maker denotes the classification error rate of the S3VM-improved, while the dashed lines with the ○, ⊙, and □ makers represent those of the three traditional S3VMs, respectively. From the figure, it can be observed that the classification accuracies of the SVM algorithms are improved by choosing helpful samples from U when using both L and U. This is clearly demonstrated in the figure where the error rates of the S3VM-improved, indicated by the ◇ marker, are lower than those of the SB-SVM, denoted using the □ symbol, for all the U_s cardinalities. From these observations, it can be determined that the proposed mechanism using the modified criterion works well with semi-supervised SVMs.

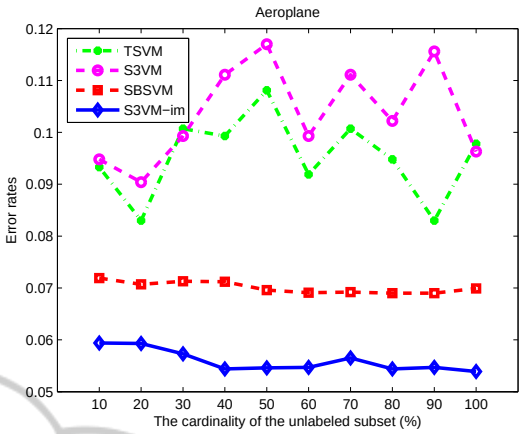

Figure 3: Comparison of the classification error rates of the two algorithms for the experimental data.

Table 3: Numerical comparison of the classification error (and standard deviation) rates (%) between the S3VM-improved and traditional algorithms for VOC'07 datasets. Here, the lowest error rate in each data set is underlined.

Datasets	S3VM-imp	TSVM	S3VM	SB-SVM
Aeroplane	5.33	8.74	10.07	7.52
	(0.44)	(0.51)	(0.93)	(0.77)
Motorbike	10.00	17.18	17.18	10.96
	(0.66)	(2.02)	(1.53)	(0.64)
Person	31.75	41.28	43.84	37.80
	(2.12)	(3.99)	(3.27)	(2.52)
Car	18.13	22.46	24.51	19.12
	(1.49)	(2.30)	(2.23)	(1.29)
Horse	10.71	17.25	20.91	12.97
	(1.05)	(3.21)	(2.32)	(1.09)

5.3 Numerical Comparison of the Error Rates

In order to further investigate the characteristics of the proposed algorithm, the experiment was repeated using different VOC'07 datasets. Table 3 presents a numerical comparison of the mean error rates and standard deviations obtained from the experiments. Here, the results in the second column were obtained using the proposed S3VM-improved algorithm where the cardinality of U_s is 10%; the results of the third, fourth, and fifth columns were obtained using the TSVM, S3VM, and SB-SVM, which were implemented using the algorithms provided in (Joachims, T., 1999b), (Chang, C. -C. and Lin, C. -J., 2011), and (Mallapragada, P. K. et al., 2009), respectively.

In addition to this result, in order to demonstrate the significant differences in the error rates between the S3VM algorithms used in the experiments, for the means (μ) and standard deviations (σ) shown in Table 3, the Student's statistical two-sample test (Huber, P. J., 1981) can be conducted. More specifically, using the t-test package, the p-value can be obtained in order to determine the significance of the difference between these algorithms. Here, the p-value represents the probability that the error rates of the S3VM-improved algorithm are generally smaller than those of the traditional S3VM algorithms.

For example, for the Motorbike dataset with $\mu_1(\sigma_1) = 0.1000(0.0066)$ for the S3VM-improved algorithm and $\mu_2(\sigma_2) = 0.1096(0.0064)$ for the SB-SVM algorithm (refer to Table 3), a p-value of 0.998 was obtained for the two algorithms. As a consequence, because $p > 0.95$ at the 5% significance level, the null hypothesis H0: $\mu_1(\sigma_1) = \mu_2(\sigma_2)$ was rejected and the alternative hypothesis H1: $\mu_1(\sigma_1) < \mu_2(\sigma_2)$ was accepted. In a similar manner, it can be observed that all Practical Image VOC'07 datasets performed better at significant levels of both 5% and 10%. From this observation, it is clear that the error rate of S3VM-improved is smaller than those of the traditional S3VM algorithms.

5.4 Comparison of the Time Complexities

Finally, the time complexity of the proposed algorithm for the VOC'07 data sets was investigated. First, Fig. 4 presents a comparison of the processing CPU-times (in seconds) obtained through repeating the above experiment ten times for the Aeroplane dataset. In the figure, the x-axis denotes the number of iterations (t) and the y-axis indicates the processing CPU-times corrupted by the two algorithms.

From Fig. 4, as mentioned in Section 3.3, it can be observed that the required time for SemiBoost is much more sensitive to the cardinalities of the training sets (L and U) and the selected data set (S) than that for the proposed algorithm. The details of the other data sets are omitted here in the interest of compactness.

Next, the processing CPU-times (in seconds) [3] of the *S3VM-imp* and *SB-SVM* methods for the VOC'07 data sets are shown in Table 4.

From the results of the table, we can see a comparison of the results obtained with the *S3VM-imp* and *SB-SVM* for the VOC'07 data sets. From these considerations, the reader should observe that the proposed philosophy of *S3VM-imp* needs less time than that of the traditional *SB-SVM* in the cases of the VOC'07 data sets.

[3]The times recorded are the times required for the MATLAB computation on a PC with a CPU speed of 2.8 GHz and RAM 4096 MB, and operating on a Window 7 Enterprise 64-bit platform.

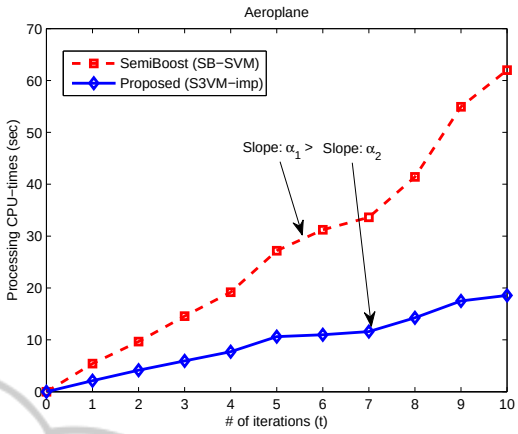

Figure 4: Comparison of the processing CPU-times (in seconds) required for the training-test computation for the experimental data set.

Table 4: Numerical comparison of the processing CPU-times (seconds) between the S3VM-improved and SB-SVM algorithms for VOC'07 datasets.

Datasets	S3VM-imp	SB-SVM
Aeroplane	18.57	62.00
Motorbike	29.18	82.92
Person	178.42	444.10
Car	58.76	159.80
Horse	30.67	86.78

6 CONCLUSIONS

In an effort to improve the classification performance of S3VM algorithms, selection criteria with which the algorithms can be implemented efficiently were investigated in this paper. S3VMs are a popular approach that attempts to improve learning performance through exploiting the whole or a subset of unlabeled data. For example, in SemiBoost, a strategy of improving the accuracy of the SVM classifier through selecting a few helpful samples from the unlabeled data has been proposed. However, the selection criterion has a weakness that is caused by the significant influence of the unlabeled data on the prediction of the labeling for the selected samples. This impact can cause errors in selecting and labeling unlabeled samples. In order to avoid this significant effect, the selection criterion was modified using the conditional

class probability estimated and the original quantities used for SemiBoost. This was motivated by an observation that the confidence levels relating to the unlabeled samples could be adjusted by subtracting the probability estimates as a penalty cost. Using the modified criterion, the confidence values relating to the labeled and unlabeled data can be balanced.

The experimental results demonstrate that the modified sampling criterion performs well with the S3VM, particularly when the impacts of the positive class and negative class are similar at the boundary. Furthermore, the results demonstrate that the classification accuracy of the proposed algorithm is superior to that of the traditional algorithms when appropriately selecting a small amount of unlabeled data. Although it has been demonstrated that S3VM can be improved using the modified criterion, many tasks remain to be improved. A significant task is the selection of an optimal, or near optimal, cardinality for the strong samples in order to further improve the classification accuracy. Furthermore, it is not yet clear which types of significant datasets are more suitable for using the selection strategy for S3VM. Finally, the proposed method has limitations in the details that support its technical reliability, and the experiments performed were limited. Future studies will address these concerns.

REFERENCES

Adankon, M. M. and Cheriet, M. (2011). Help-training for semi-supervised support vector machines. In *Pattern Recognition*, volume 44, pages 2946–2957.

Ben-David, S., Lu, T., and Pal, D. (2008). Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proc. the 22th Ann. Conf. Computational Learning Theory (COLT08)*, pages 33–44, Helsinki, Finland.

Bennett, K. P. and Demiriz, A. (1998). Semi-supervised support vector machines. In *Proc. Neural Information Processing Systems*, pages 368–374.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proc. the 11th Ann. Conf. Computational Learning Theory (COLT98)*, pages 92–100, Madison, WI.

Chakraborty, S. (2011). Bayesian semi-supervised learning with support vector machine. In *Statistical Methodology*, volume 8, pages 68–82.

Chang, C. -C. and Lin, C. -J. (2011). LIBSVM : a library for support vector machines. In *ACM Trans. on Intelligent Systems and Technology*, volume 2, pages 1–27.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press, Cambridge, MA.

Dagan, I. and Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. In A. Prieditis, S. J. Russell, editor, *Proc. Int'l Conf. on Machine Learning*, pages 150–157, Tahoe City, CA.

Du, J., Ling, C. X., and Zhou, Z. -H. (2011). When does cotraining work in real data? In *IEEE Trans. on Knowledge and Data Eng.*, volume 23, pages 788–799.

Duin,R. P. W., Juszczak, P., de Ridder, D., Paclik, P., Pekalska, E., and Tax, D. M. J. (2004). *PRTools 4: a Matlab Toolbox for Pattern Recognition*. Delft University of Technology, The Netherlands.

Everingham, M., Van Gool, L., William, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Goldberg, A. B. (2010). *New Directions in Semi-Supervised Learning*. University of Wisconsin - Madison, Madison, WI.

Goldberg, A. B., Zhu, X., Singh, A., Zhu, Z., and Nowak, R. (2009). Multi-manifold semi-supervised learning. In D. van Dyk, M. Welling, editor, *Proc. the 12th Int'l Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 99–106, Clearwater, FL.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York, NY.

Jiang, Z., Zhang, S., and Zeng, J. (2013). A hybrid generative/discriminative method for semi-supervised classification. In *Knowledge-Based System*, volume 37, pages 137–145.

Joachims, T. (1999a). Making large-Scale SVM Learning Practical. In B. Sch?lkopf, C. Burges, A. Smola, editor, *Advances in Kernel Methods - Support Vector Learning*, pages 41–56, Cambridge, MA. The MIT Press.

Joachims, T. (1999b). Transductive inference for text classification using support vector machines. In *Proc. the 16th Int'l Conf. on Machine Learning*, pages 200–209, San Francisco, CA. Morgan Kaufmann.

Kuo, H. -K. J. and Goel, V. (2005). Active learning with minimum expected error for spoken language understanding. In *Proc. the 9th Euro. Conf. on Speech Communication and Technology*, pages 437–440, Lisbon. Interspeech.

Le, T. -B. and Kim, S. -W. (2012). On improving semi-supervised MarginBoost incrementally using strong unlabeled data. In P. L. Carmona, J. S. Sánchez, and A. Fred, editor, *Proc. the 1st Int'l Conf. Pattern Recognition Applications and Methods (ICPRAM 2012)*, pages 265–268, Vilamoura-Algarve, Portugal.

Leng, Y., Xu, X., and Qi, G. (2013). Combining active learning and semi-supervised learning to construct SVM classifier. In *Knowledge-Based Systems*, volume 44, pages 121–131.

Li, Y. -F. and Zhou, Z. -H. (2011). Improving semi-supervised support vector machines through unlabeled instances selection. In *Proc. the 25th AAAI Conf. on Artificial Intelligence (AAAI'11)*, pages 386–391, San Francisco, CA.

Lu, T. (2009). *Fundamental Limitations of Semi-Supervised Learning*. University of Waterloo, Waterloo, Canada.

Mallapragada, P. K., Jin, R., Jain, A. K., and Liu, Y. (2009). SemiBoost: Boosting for semi-supervised learning. In

IEEE Trans. Pattern Anal. and Machine Intell., volume 31, pages 2000–2014.

McClosky, D., Charniak, E., and Johnson, M. (2008). When is Self-Training Effective for Parsing? In *Proc. the 22nd Int'l Conf. Computational Linguistics (Coling 2008)*, pages 561–568, Manchester, UK.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. In *Machine Learning*, volume 39, pages 103–134.

Riccardi, G. and Hakkani-Tur, D. (2005). Active learning: theory and applications to automatic speech recognition. In *IEEE Trans. on Speech and Audio Processing*, volume 13, pages 504–511.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *Proc. the 7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)*, pages 29–36, Breckenridge, CO.

Singh, A., Nowak, R., and Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't. In T. Matsuyama, C. Cipolla, et al., editor, *Advances in Neural Information Processing Systems (NIPS)*, pages 1513–1520, London. The MIT Press.

Sun, S., Hussain, Z., and Shawe-Taylor, J. (2014). Manifold-preserving graph reduction for sparse semi-supervised learning. In *Neurocomputing*, volume 124, pages 13–21.

Sun, S. and Shawe-Taylor, J. (2010). Sparse semi-supervised learning using conjugate functions. In *Journal of Mach. Learn. Res.*, volume 11, pages 2423–2455.

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data (English translation 1982, Russian version 1979.)*. Springer, New York.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vapnik, V. and Chervonenkis, A. I. (1974). *Theory of Pattern Recognition*. Nauka, Moscow.

Vedaldi, A. and Zisserman, A. (2011). Image Classification Practical, 2011.

Zhu, X. (2006). *Semi-Supervised Learning Literature Survey*. University of Wisconsin - Madison, Madison, WI.