On Selecting Useful Unlabeled Data Using Multi-view Learning Techniques*

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

Keywords: Semi-supervised Learning, Selecting Unlabeled Data, Multi-view Learning Techniques.

Abstract: In a semi-supervised learning approach, using a selection strategy, strongly discriminative examples are first selected from unlabeled data and then, together with labeled data, utilized for training a (supervised) classifier. This paper investigates a new selection strategy for the case when the data are composed of different multiple views: first, multiple views of the data are derived independently; second, each of the views are used for measuring corresponding confidences with which examples to be selected are evaluated; third, all the confidence levels measured from the multiple views are used as a weighted average for deriving a target confidence; this selecting-and-training is repeated for a predefined number of iterations. The experimental results, obtained using synthetic and real-life benchmark data, demonstrate that the proposed mechanism can compensate for the shortcomings of the traditional strategies. In particular, the results demonstrate that when the data is appropriately decomposed into multiple views, the strategy can achieve further improved results in terms of the classification accuracy.

1 INTRODUCTION

In semi-supervised learning (SSL) approaches (Zhu, X. and Goldberg, A. B., 2009), a large amount of unlabeled data (U), together with labeled data (L), is used to build better classifiers. 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. Therefore, in order to select a small amount of useful unlabeled data, various sampling (and selecting) techniques have been proposed in the literature, including the self-training (Yarowsky, 1995), co-training (Blum, A. and Mitchell, T., 1998), confidence-based approaches (Le, T. -B. and Kim, S. -W., 2014; Mallapragada, P. K. et al., 2009), and other approaches used in active learning (AL) algorithms (Dagan, I. and Engelson, S. P., 1995; Reitmaier, T. and Sick, B., 2013). However, in AL, selected instances are useful when they are labeled, thus it is required to query their true class label from a human annotator. From this point of view, the approaches for SSL and AL algorithms are different.

In SemiBoost (Mallapragada, P. K. et al., 2009), for example, the confidence level of all $x_i \in U$ examples is first computed based on the prediction made by an ensemble classifier and the similarity among the examples of $L \cup U$. Then, a few examples with higher confidence are selected to retrain the ensemble classifier together with L. This selecting-and-training step is repeated for a predefined number of iterations or until a termination criterion is met. More specifically, the confidence level is computed using two quantities, i.e. $p_i (= p(x_i))$ and $q_i (= q(x_i))$, which are measured using the pairwise similarity between $x_i \in U$ and other L and U examples. 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 useful examples.

In order to address this problem, a modified criterion that minimizes the errors in estimating the labels of U examples has been considered (Le, T. -B. and Kim, S. -W., 2014). In the modified criterion, for $x_i \in U$ that are near the boundary between the positive class of L (L^+) and the negative class of L (L^-), the confidence levels (and predicted pseudo labels) are measured using estimates of the class-conditional probabilities (P_E) as well as the quantities of p_i and q_i. However, sometimes the modified criterion devel-

*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 Useful Unlabeled Data Using Multi-view Learning Techniques.
DOI: 10.5220/0005171301570164
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 157-164
ISBN: 978-989-758-076-5

ops a weakness. In general, the reason for this can be explained as follows: for the modified criterion, the confidence level, which is denoted using a $\rho 1(x_i)$ symbol, is determined with four terms of the L^+, L^-, U concerned terms and P_E included as a *certainty level*. Thus, when L^- and L^+ terms are similar, as a consequence $\rho 1(x_i)$ depends on U and P_E terms only. Therefore, when L examples are highly overlapped, P_E cannot work well and in turn, $\rho 1(x_i)$, which completely depends on the deeply overlapped U examples, cannot work as well. In order to address this issue, a selection strategy utilizing multi-view learning techniques is investigated in this paper.

Currently, multi-view data is common in a wide variety of applications (Kumar, A. and Daume III, H., 2011). This is the motivation for an approach of selecting useful unlabeled data based on the multi-view setting in this paper: first, a partition of the attributes of the data is found using either a similarity measure between pair-wise examples or a feature selection function; corresponding views are derived independently; each of the views is used for measuring its confidence; finally, all the confidences measured from the views are used as a weighted average for deriving a target confidence. The central idea to the proposed strategy is that the confidence obtained from one view could be helpful for deriving another confidence from another view. The strategy is empirically compared with some traditional methods using synthetic and real-world data.

The main contribution of this paper is the demonstration that the classification accuracy of the supervised / semi-supervised classifiers can be improved using the multi-view based selection strategy. In particular, it is demonstrated that when the data is *sufficiently* decomposed into multiple views, the classification accuracy is significantly improved. The remainder of the paper is organized as follows: in Section 2, a brief introduction to the selection criteria used in SemiBoost and its modified version is provided; in Section 3, a method of utilizing the multi-view based criterion is presented; in Section 4, an illustrative example for comparing the selection criteria and the experimental results obtained using the real-life datasets are presented; in Section 5, the concluding remarks are presented.

2 RELATED WORK

2.1 Sampling in SemiBoost

The goal of SemiBoost is to iteratively improve the performance of a supervised learning algorithm ($\mathcal{A}$) by using U and pairwise similarity. In order to follow the boosting idea, SemiBoost optimizes performance through minimizing an objective loss function defined as Proposition 2 in (Mallapragada, P. K. et al., 2009)), where $h_i(=h(x_i))$ is the base classifier learned by $\mathcal{A}$ at the iteration, α is the weight for combining h_i's, and

$$\begin{aligned} p_i &= \sum_{j=1}^{n_l} S^{ul}_{i,j} e^{-2H_i} \delta(y_j,1) + \frac{C}{2} \sum_{j=1}^{n_u} S^{uu}_{i,j} e^{H_j - H_i}, \\ q_i &= \sum_{j=1}^{n_l} S^{ul}_{i,j} e^{2H_i} \delta(y_j,-1) + \frac{C}{2} \sum_{j=1}^{n_u} S^{uu}_{i,j} e^{H_i - H_j}. \end{aligned} \quad (1)$$

Here, $H_i(=H(x_i))$ denotes the final combined classifier and S denotes the pairwise similarity: $S(i,j) = exp(-\|x_i - x_j\|_2^2/\sigma^2)$ for all x_i and x_j of the training set, where σ is the scale parameter controlling the spread of the function. In addition, S^{lu} (and S^{uu}) denotes the $n_l \times n_u$ (and $n_u \times n_u$) submatrix of S, where $n_l = |L|$ and $n_u = |U|$. Also, S^{ul} and S^{ll} can be defined correspondingly; the constant C, which is computed using $C = |L|/|U| = n_l/n_u$, 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 and negative class, respectively. That is, p_i and q_i can be used to guide the selection of U examples at each iteration using the confidence measurement $|p_i - q_i|$, as well as to assign the pseudo class label $sign(p_i - q_i)$. From (1), a selection criterion can be formulated as follows:

$$\begin{aligned} p_i - q_i = &\left(e^{-2H_i} \sum_{x_j \in L^+} S^{ul}_{i,j} \right) - \left(e^{2H_i} \sum_{x_j \in L^-} S^{ul}_{i,j} \right) \\ &+ \left(\frac{C}{2} \sum_{x_j \in U} S^{uu}_{i,j} (e^{H_j - H_i} - e^{H_i - H_j}) \right), \end{aligned} \quad (2)$$

where $L^+ \equiv \{(x_i,y_i)|y_i = +1, i = 1,\cdots,n_l^+\}$ and $L^- \equiv \{(x_i,y_i)|y_i = -1, i = 1,\cdots,n_l^-\}$ are the L examples in class $\{+1\}$ and class $\{-1\}$, respectively. In (2), by substituting the three corresponding summations with X_i^+, X_i^-, and X_i^u symbols, the criterion can be represented as: $p_i - q_i = X_i^+ - X_i^- + X_i^u$.

However, providing more data is not always beneficial. If the value obtained using the third term in (2) is very large or X_i^+ is nearly equal to X_i^-, (2) 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 examples. That is, the L examples do not affect the estimation of x_i label; therefore, the estimated label is unsafe and untrustworthy.

2.2 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.

In order to avoid this, the criterion of (2) can be improved through balancing the three terms in (2), i.e. X_i^+, X_i^-, and X_i^u. Using the probability estimates as a penalty cost, the criterion of (2) can be modified as follows:

$$\rho 1(x_i) = \left| X_i^+ - X_i^- + X_i^u - (1 - P_E(x_i)) \right|, \quad (3)$$

where $P_E(x_i)$ denotes the class posterior probability of instance of x_i (i.e. a *certainty level* and $1 - P_E(x_i)$ corresponds to the percentage of mistakes when labeling x_i.

However, it should be noted that sometimes using (3) leads to a problematic situation. When all $x_i \in U$ are well-separated and correctly distributed, $P_E(x_i)$ can be measured successfully and, in turn, the criterion of (3) works well. However, for the case where the $L \cup U$ examples are composed of (different) multiple views (i.e., feature vectors of the data consist of different attribute subsets) and, especially, they are near the boundary between two classes, $\rho 1(x_i)$ may not work for selecting good examples from U. Generally, the reason for this can be explained as follows. $\rho 1(x_i)$ is computed from distance measurements between pair-wise examples of $L \cup U$. Under $X_i^+ \approx X_i^-$, computation of $\rho 1(x_i)$ is completely determined from X_i^u and $P_E(x_i)$. As a consequence, deeply overlapped distribution of U examples leads to a problematic situation, where $P_E(x_i)$ cannot work well and, in turn, $\rho 1$ cannot work as well.

3 PROPOSED METHOD

3.1 Newly Modified Criterion

Assume that a data set $X = \{x_1, \cdots, x_n\}$, $x_i \in \mathbb{R}^d$ consists of two views with their respective feature partitions: $X_1 = \{x_{11}, \cdots, x_{1n}\}$, $x_{1i} \in \mathbb{R}^{d_1}$, and $X_2 = \{x_{21}, \cdots, x_{2n}\}$, $x_{2i} \in \mathbb{R}^{d_2}$, where $d = d_1 + d_2$. That is, they are two attribute subsets describing the data X. In the multi-view learning (de Sa, V., 1994), two classifiers, h_1 and h_2, are trained using L, both on their respective views; they operate on different views of an example: h_1 is based on L_1, h_2 is based on L_2. The two classifiers are then evaluated using U; after classifying the remaining data of U, with h_1 and h_2 separately, the examples (i.e., U_s) on which it is the most confident are removed from U and added to L for the next iteration, i.e. $\{(x_i, h_1(x_i))\}$ to L_2 and $\{(x_i, h_2(x_i))\}$ to L_1. Both classifiers are now retrained on the expanded L_1 and L_2 sets and the procedure is repeated until some stopping criteria is met.

In order to select useful unlabeled data, the multi-view learning technique can be utilized as follows: first, each of the related data sets is divided into two different views (i.e., L_k, U_k, and S_k, where $k = 1, 2$) using a feature selection scheme; next, as in (2), for all $x_{ki} \in U_k$, after designing h_k per view, the confidence levels of x_{ki}, using $|p_{ki} - q_{ki}|$, and the classification error rates $\varepsilon(h_k)$, using all L_j, $(j \neq k)$, are computed; finally, for all $x_i \in U$, using $\varepsilon(h_k)$ as a weight, its confidence level (which is newly denoted using a $\rho 2(x_i)$ symbol) can be measured as follows:

$$\rho 2(x_i) = \sum_{k=1}^{K} \frac{1}{\varepsilon\ (h_k)} \left| p_{ki} - q_{ki} \right|. \quad (4)$$

Using (4) as a selection criterion, a sampling function, named Multi-view Sampling, can be summarized as follows.

Algorithm 1: Multi-view Sampling.

Input: Labeled data (L), unlabeled data (U), view # of attributes ($K = 2$), and α (% of selected U over U).

Output: Selected unlabeled data (U_s).

Procedure: Perform the following steps.

1. Decompose $L \in \mathbb{R}^d$ into L_k, $(k = 1, \cdots, K)$, views first (refer to (1)), where $L_k \in \mathbb{R}^{d_k}$ and $d = \sum_{k=1}^{K} d_k$; then, train a classifier (h_k) per view using corresponding data L_k.
2. First, compute a mapping function (w_k) to transform L to L_k, i.e. $L_k \leftarrow L \cdot w_k$; second, using w_k, compute $U_k \leftarrow U \cdot w_k$; third, compute the error rates, $\varepsilon(h_k)$, using all L_j, $(j \neq k)$.
3. For all $x_{ki} \in U_k$, first, compute confidence levels (ρ_{ki}) per view using p_i, q_i, and h_k; then, average the levels as $\rho 2_i \leftarrow \sum_{k=1}^{K} \frac{1}{\varepsilon(h_k)} \rho_{ki}$, and sort them in a decreasing order on key $\{|\rho 2_i|\}$.
4. Choose U_s from top $\alpha(\%)$ of U and estimate predicted labels of $x_j \in U_s$ using $sign(\rho 2_j)$.

End Algorithm

3.2 Proposed Algorithm

The proposed learning algorithm initially predicts the labels of all U examples using a classifier (SVM, for example) trained with L only. After initializing the related parameters, e.g. the kernel function and its related conditions, the confidence levels of all U examples (i.e., $\{|\rho 2(x_i)|\}_{i=1}^{n_u}$) are calculated using (3). Then, $\{|\rho 2(x_i)|\}_{i=1}^{n_u}$ is sorted in descending order. After selecting the examples ranked with the highest confidence levels, adding them to L creates an expanded training set (T).

Finally, the above selecting-and-training step is 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 examples (i.e. $T = L \cup U_s$), constitutes the final classifier for the problem.

Based on this brief explanation, a learning algorithm, using the sampling function in **Algorithm 1**, is summarized as follows: the labeled and unlabeled data (L and U), cardinality of U_s, number of iterations (e.g. $J = 10$ and $K = 2$), 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 2: Proposed Learning Algorithm.
Input: Labeled data ($L$) and unlabeled data ($U$). Output: Final classifier ($H$). Method: Initialization: Set the parameters: $J$, $K$, $\alpha^{(0)}$, $\Delta^{(0)}$; train a classifier $H$ (and its error rate $\varepsilon(H)$) using $L$.
Procedure: Repeat the following steps while increasing $j$ from 1 to $J$ in increments of 1.
  1. Obtain $U_s^{(j)}$ (and their predicted labels) from $U$ by invoking the Multi-view sampling function with training data $T$ (or $L$ if $j = 1$), $U$, $K$, $\alpha^{(j)}$.
  2. Update the training data $T$ using $U_s^{(j)}$, i.e. $T \leftarrow L \cup U_s^{(j)}$, and train a classifier ($h_j$) (and its classification error rate, named $\varepsilon(h_j)$) through $T$.
  3. If $\varepsilon(h_j) \leq \varepsilon(H)$, then keep $h_j$ as the best classifier, i.e. $H \leftarrow h_j$ and $\varepsilon(H) \leftarrow \varepsilon(h_j)$.
  4. $\alpha^{(j+1)} \leftarrow \alpha^{(j)} + \Delta^{(j)}$.
End Algorithm
```

4 EXPERIMENTAL RESULTS

4.1 Synthetic Data

First, two 2-dimensional, two-class datasets having different Gaussian distributions were generated. Then, the two 2-dimensional datasets were concatenated into a 4-dimensional, two-class dataset.

For the synthetic data, an experiment was conducted as follows: first, two confidence values were computed for all $x_i \in U$ using the three criteria in (2), (3), and (4), i.e. $|p_i - q_i|$, $|\rho 1(x_i)|$, and $|\rho 2(x_i)|$; second, a subset of U, i.e. U_s (the 10% cardinality of U), was selected referring to the confidence values. These two steps were repeated *ten* times. Fig. 1 presents a comparison of three selections obtained with the experiment. Here, the data of $(5,5)$ objects of the two classes were generated ten times as L. For each generation of the L data, the data of $(500,500)$ objects were generated ten times again as U.

From the figure, it can be observed that the capability of the proposed strategy for selecting useful unlabeled data for discrimination is generally improved. This is clearly demonstrated in the differences between Fig. 1 (a), (b), and (c) by the number of selected objects and their geometrical structures. More specifically, for the boundary regions between the positive and negative classes in Fig. 1 (a), the number of the selected objects but overlapped of the multi-view strategy is smaller than that of the original and modified strategies in Fig. 1 (b) and (c). In contrast, for all the three criteria, there is no object being selected but overlapped. The selected objects are all appropriately chosen. From this observation, it should be noted that the discriminative power of the multi-view strategy might be better than that of the original SemiBoost and its modified strategies.

In order to investigate this further, another experiment was conducted on labeling the (selected) unlabeled data using the three selection strategies. That is, a verification of the pseudo-labels for each $x_i \in U$ was predicted using the three criteria. The experiment was undertaken as follows: first, as in the above experiment, a subset (U_s) was selected from U after computing the confidence values; second, three pseudo-labels of all $x_i \in U_s$ were predicted using the three techniques in (2), (3), and (4), i.e. $sign(p_i - q_i)$, $sign(\rho 1(x_i))$, and $sign(\rho 2(x_i))$; third, the predicted pseudo-labels were compared with their true labels ($y_{U_s} \in \{+1,-1\}$); finally, based on the comparison, the number of wrongly predicted objects was counted. Based on this count, in order to clearly compare the selection criteria, a wrong-prediction rate ($\varepsilon_{Prediction}$) was defined as: $\varepsilon_{Prediction} = \frac{Failed\ Prediction\ \#}{Total\ Prediction\ \#}$.

From Fig. 1 (b), a high value of $\varepsilon_{Prediction} (= 46/100)$ was obtained using the original criterion. In contrast, from Fig. 1 (a) and (b), a very small value of $\varepsilon_{Prediction} (= 0/100)$ was obtained using the two modified criteria, i.e. Multi-view and Modified. As mentioned previously, although the wrong-prediction rates of the two criteria are the same, among the selected objects, the number of overlapped objects of

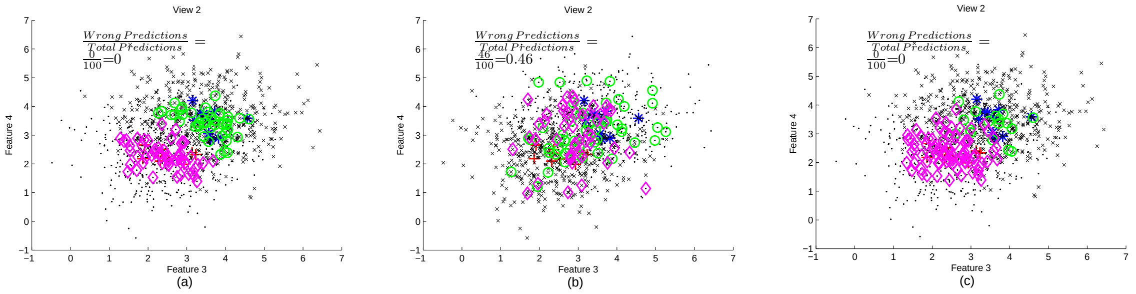

Figure 1: Plots comparing the selected objects with the three selection criteria: (a) the multi-view criterion, (b) the original criterion, and (c) its modified criterion for the four-dimensional, two-class synthetic data. The results obtained are partially displayed in $[x_3, x_4]$ only (where $[x_1, x_2]$'s are almost the same and omitted here); the objects in the positive and negative classes are denoted by '$+$' (in red) and '$*$' (in blue) symbols, respectively; the selected objects from the two classes are marked with '$\diamond$' (in pink) and '$\circ$' (in green) symbols, respectively; the unlabeled data are indicated using a '$\cdot$' symbol.

Table 1: Comparison of characteristics observed with two decomposition methods, Random and `featseli`.

Division methods	Attributes of two views		Prediction	Classification
	View 1	View 2	$\varepsilon_{Prediction}$	$\varepsilon_{Classification}$
Random	$[x_2, x_3]$	$[x_1, x_4]$	0 / 100	8/100
`featseli`	$[x_1, x_2]$	$[x_3, x_4]$	0 / 100	5/100

the latter criterion is larger than that of the former. From this observation, it should be noted again that the discriminative power of the newly proposed criterion might be better than that of the original criterion and its modified version.

Finally, it should also be mentioned that the L and U data were *artificially* divided into two views of $[x_1, x_2]$ and $[x_3, x_4]$ attributes. However, a crucial aspect of the multi-view based selection criterion is that the two classifiers are trained on *class-conditionally independent* and *sufficiently* different views of the data set (Zhu, X. and Goldberg, A. B., 2009). In order to investigate the underlying reason for this, attributes of the two views were selected using two decomposition methods: (1) randomly decomposition as a baseline method and (2) decomposition based on a traditional feature selection method, such as `featseli` (individual feature selection) package of PRTools [2]. Table 1 presents a comparison of characteristics achieved with the two decomposition methods for the multi-views.

From Table 1, it should be noted that the classification accuracy could generally be improved when using the U_s obtained using `featseli`, while the wrong-prediction rates of them are the same. However, Random method did not work satisfactorily with selecting the number of attributes of the two views. In `featseli`, a 4-dimensional vector is decomposed into two 2-dimensional vectors, $[x_1, x_2]$ and $[x_3, x_4]$, correctly and naturally. Meanwhile, in Random, the decomposition was performed unnaturally. From these considerations, the reader should observe that the multi-view based criterion can select helpful data from U more efficiently, rather than selecting them using the original SemiBoost and modified criteria.

4.2 USPS and UCI Data

In order to further investigate the run-time characteristics of the multi-view based selection, comparisons were accomplished through performing experiments on real-life datasets cited from USPS Handwritten Digits (Hastie, T. et al., 2001) and the UCI Machine Learning Repository (Asuncion, A. and Newman, D. J., 2007).

The first experiment focused on a simple but clearly *multi-viewed* classification problem. In order to achieve this goal, as was done in (Chen, M. et al., 2011), the USPS dataset was reconstructed to a binary data (named USPS2). Here, each instance in the USPS2 was composed of digits sampled from the USPS handwritten digits set; for the positive class (Class +), *View 1* was uniformly picked from the set of ones and twos and *View 2* was picked from the set of fives or sixes; for the negative class (Class -), *View 1* was a three or four and *View 2* a seven or eight. As a result, USPS2 was a *two-views* dataset of 512-dimensional, two-class, and 2000 objects.

The second experiment was done with the benchmark datasets of the UCI repository. The specific characteristics (dimension # / object # / class #) of the UCI datasets are: Australian (14 / 2 / 690), Credit A (14 / 2 / 653), Ecoli (7 / 8 / 336), Glass (9 / 6 / 214), Heart (13 / 2 / 270), Pima (5 / 2 / 768), Quality (13 / 7 / 4898), Segment (19 / 7 / 2310), Vehicle (18 / 4 / 846), and Vowel (10 / 11 / 528).

In two consecutive experiments, each dataset was divided into three subsets: a labeled training set (L),

[2] http://prtools.org/

Table 2: Classification error rates (%) between the Multi-view and traditional strategies for the USPS2 dataset.

Datasets	Multi-view	SemiBoost	Modified	S3VM-us
USPS2	0.42	20.87	28.03	27.27

labeled test set (T_e), and unlabeled data set (U), 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. For simplicity, in the S3VM classifier (which was implemented using the algorithm provided in (Chang, C. -C. and Lin, C. -J., 2011)), the two constants, C^* and C, were set to 0.1 and 100, respectively. 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 learning algorithm, in which a S3VM was trained with L and U_s selected using the proposed criterion, was compared with three types of traditional algorithms: SemiBoost, Modified, and S3VM-us. In these S3VM algorithms, U_s's were selected using the selection criteria provided in (Mallapragada, P. K. et al., 2009), (Le, T. -B. and Kim, S. -W., 2014), and (Li, Y. -F. and Zhou, Z. -H., 2011). Also, the cardinality of U_s's was fixed as 10% of U.

4.3 Experimental Results obtained with USPS2

In order to investigate the characteristics of the proposed selection strategy, classification was performed using the *two-view* USPS2 data, which has been synthetically reconstructed from the original dataset.

Table 2 presents a numerical comparison of the mean error rates (and standard deviations) (%) obtained with the S3VM classifiers. Here, the results in the second column (Multi-view) were obtained using the proposed learning algorithm (Algorithm 2). The results of the third, fourth, and fifth columns were obtained using the selection strategies of SemiBoost, Modified, and S3VM-us, respectively.

From Table 2, the reader should observe that the classification accuracy of S3VM could be significantly improved when using the U_s selected using the multi-view based criterion. From this observation, the rationale that the multi-view based criterion (Multi-view) developed in the present work, rather than the distance and/or density based criteria (i.e., SemiBoost, Modified, and S3VM-us), has been employed as a selection strategy is clear.

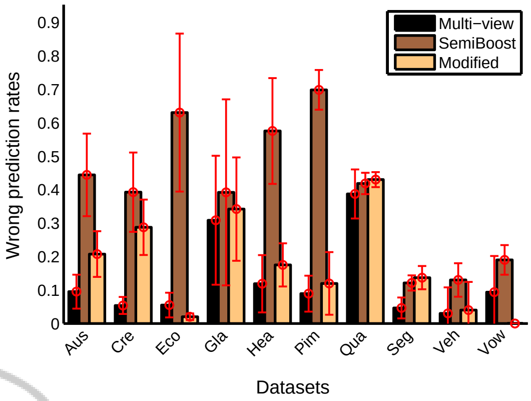

Figure 2: Comparison of the wrong-prediction rates among the three selection criteria for the experimental data. Here, the datasets are represented with three letter acronym.

4.4 Wrong-prediction Rates: UCI Datasets

Prior to presenting the classification accuracies, the three criteria, i.e. newly proposed Multi-view, original SemiBoost, and Modified, were compared. First, the following question was investigated: *does the newly proposed selection criterion perform better than the traditional criteria?* To answer this question, an experiment was conducted on labeling the unlabeled data using the three selection criteria: a verification of the two predicted labels for each $x_i \in U$ using two criteria. As was done for the synthetic data, the experiment was undertaken as follows: first, a subset from U (i.e., U_s) was selected using one of the three criteria; second, the three labels of all $x_i \in U_s$ were predicted using the three techniques in (2), (3), and (4), i.e. $sign(p_i - q_i)$, $sign(\rho1(x_i))$, and $sign(\rho2(x_i))$ were compared with their true labels ($y_{U_s} \in \{+1,-1\}$); the above two steps were repeated ten times without increasing the cardinality of U_s, which means that, in Algorithm 2, for all j, $\Delta^{(j)} = 0$. Fig. 2 presents a comparison of the ten values obtained through performing the above experiment for the UCI datasets. In the figure, the x-axis denotes UCI datasets and the y-axis indicates the incorrect prediction rates obtained using the three criteria.

From the figure, it can be observed that the prediction capabilities of the three criteria differ from each other; in general, the capability of the proposed strategy, Multi-view, appears better than that of the traditional strategies, SemiBoost and Modified. This can be clearly demonstrated by comparing the wrong-prediction rates in the figure. For almost all the datasets, the lowest rate was observed with Multi-view, rather than SemiBoost and Modified.

In addition to this observation, for Ecoli and Vowel datasets, the prediction capability of Modified seems to be superior to that of Multi-view. This means that the proposed strategy does not work well with the two datasets. In order to discover the reason for this poor performance, data complexities (Ho, T. K. and Basu, M., 2002) were investigated as follows: first, two subsets of the UCI datasets, i.e. UCI1: (Cre, Aus, Hea) and UCI2: (Eco, Vow) were considered; next, for the datasets of the two subsets, six kinds of complexities, namely $F3$ (individual feature efficiency), $F4$ (collective feature efficiency), $N1$ (the fraction of points on the class boundary), $N2$ (ratio of average intra/inter class nearest neighbor distance), $N3$ (the leave-one-out error rate of the one-nearest neighbor classifier, 1NN), and $N4$ (the nonlinearity of 1NN) were measured. Details of the measurement are omitted here in the interest of space, but observations obtained can be summarized as follows: for the N measures, each value of UCI1 dataset is larger than that of UCI2 dataset, whereas for the F measures, the result is the opposite.

In review, as can be seen from Fig. 2, for a few datasets, such as Ecoli and Vowel, the prediction capability of the proposed strategy seems to be inferior to that of the traditional strategies, which means that Multi-view does not work with certain kinds of datasets. In order to figure out why this is, the data complexities of the F and N measures were considered. From this measurement, it has been demonstrated that the data sets, with which the new criterion does not work, seem to be composed of *similar views*, not different. From this observation, the reason that the multi-view based criterion, rather than the distance and/or density based criteria, has been employed as a selection strategy, is clear again.

4.5 Experimental Results obtained with UCI

In order to further investigate the characteristics of the proposed selection strategy and to determine which types of datasets are more suitable for it, classification was performed using the proposed and three traditional learning algorithms for the UCI datasets. In particular, S3VM classifiers (Chang, C. -C. and Lin, C. -J., 2011) were used. Here, the U_s has been selected using four different criteria: Mult-view, SemiBoost, Modified, and S3VM-us.

Table 3 presents a numerical comparison of the mean error rates (and standard deviations) (%) obtained with the S3VM classifiers. Here, the results in the second column (Multi-view) were obtained using the proposed learning algorithm (Algorithm 2). The results of the third, fourth, and fifth columns were obtained using the selection strategies of the original SemiBoost algorithm (Mallapragada, P. K. et al., 2009), the modified SemiBoost algorithm (Le, T. -B. and Kim, S. -W., 2014), and the S3VM-us algorithm (Li, Y. -F. and Zhou, Z. -H., 2011) respectively. For all algorithms, the cardinality of U_s is 10% (i.e., $\alpha^{(j)} = 10$ for all j).

Table 3: Classification error rates (%) between the Multi-view and traditional algorithms for the UCI datasets. Here, the lowest error rate in each data set is underlined.

Datasets	Multi-view	SemiBoost	Modified	S3VM-us
Australian	27.23	39.05	36.06	32.41
Credit A	29.54	40.31	35.46	35.23
Ecoli	3.33	3.64	3.18	3.94
Glass	38.10	35.95	36.67	35.00
Heart	40.37	44.26	43.52	44.63
Pima	29.67	34.38	34.71	34.05
Quality	43.06	44.27	43.66	44.26
Segment	7.64	17.88	13.70	14.29
Vehicle	12.32	23.45	20.12	22.44
Vowel	5.05	12.76	4.38	5.71

From Table 3, it should be observed that the classification accuracy of S3VM could generally be improved when using the U_s through the Multi-view criterion. For example, consider the results for the Australian dataset. For the dataset ($d = 14$), the lowest error rate (27.23 %) was obtained using Multi-view. However, as observed previously, the proposed criterion did not work satisfactorily with certain kinds of datasets, such as Ecoli, Glass, and Vowel.

Although it is hard to quantitatively compare the four criteria, to render this comparative evaluation more complete, we counted the numbers of the underlined error rates, obtained with the ten UCI datasets. In summary, the numbers of the underlined error rates for the four columns of Multi-view, SemiBoost, Modified, and S3VM-us are, respectively, 7, 0, 2, and 1. From this, it can be observed that Multi-view, albeit not always, generally works better than the others in terms of the classification accuracy.

In addition to this simple comparison, in order to demonstrate the significant differences in the error rates among the selection criteria used in the experiments, for the means (μ) and standard deviations (σ) shown in Table 3, the Student's statistical two-sample test 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 the Multi-view and Modified criteria. Here, the p-value represents the probability that the error rates of the former are generally smaller than those of the latter. More details of this observation are omitted here, but will be reported in the journal version.

5 CONCLUSIONS

In an effort to improve the classification performance of SSL based models, selection criteria with which the models can be implemented efficiently were investigated in this paper. With regard to this concern, various approaches have been proposed in the literature. Recently, for example, a selection strategy of improving the accuracy of a SemiBoost classifier has been proposed. However, the criterion has a weakness, which is caused by the significant influence of the unlabeled data when predicting the labels of the selected examples. In order to avoid this weakness, a selection strategy utilizing the multi-view learning techniques was studied and compared with the original strategy used for SemiBoost and its variants.

Experiments have been performed using synthetic and real-life benchmark data, where the data was uniformly decomposed into two-views through a traditional feature selection method was applied for extracting the views. The experimental results obtained demonstrated that the proposed mechanism can compensate for the shortcomings of the traditional strategies. In particular, the results demonstrated that when the data is decomposed into multiple views efficiently, i.e. the data has multiple different views as its real nature, the strategy can achieve further improved results in terms of prediction/classification accuracy.

Although it has been demonstrated that the accuracy can be improved using the proposed strategy, more tasks should be carried out. A significant task is the decomposition of the multiple views from the labeled and unlabeled data to measure correct confidence labels of the unlabeled data. Furthermore, it is not yet clear which types of datasets are more suitable for using this multi-view based selection strategy for SSL. Finally, the proposed method has limitations in the details that support its technical reliability, and the experiments performed were limited.

REFERENCES

Asuncion, A. and Newman, D. J. (2007). UCI Machine Learning Repository. Irvine, CA. University of California, School of Information and Computer Science.

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.

Chang, C. -C. and Lin, C. -J. (2011). LIBSVM : a library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2(3):1–27.

Chen, M., Weinberger, K. Q., and Chen, Q. (2011). Automatic feature decomposition for single view co-training. In *Proc. of the 28 th Int'l Conf. Machine Learning (ICML-11)*, pages 953–960, Bellevue, Washington, USA.

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.

de Sa, V. (1994). Learning classification with unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 6, pages 112–119.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Ho, T. K. and Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. and Machine Intell.*, 24(3):289–300.

Kumar, A. and Daume III, H. (2011). A co-training approach for multi-view spectral clustering. In Getoor, L. and Scheffer, T., editors, *Proc. of the 28th Int'l Conf. on Machine Learning (ICML-11)*, ICML '11, pages 393–400, New York, NY, USA. ACM.

Le, T. -B. and Kim, S. -W. (2014). On selecting helpful unlabeled data for improving semi-supervised support vector machines. In A. Fred, M. de Marsico, and A. Tabbone, editor, *Proc. the 3rd Int'l Conf. Pattern Recognition Applications and Methods (ICPRAM 2014)*, pages 48–59, Angers, France.

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.

Mallapragada, P. K., Jin, R., Jain, A. K., and Liu, Y. (2009). SemiBoost: Boosting for semi-supervised learning. *IEEE Trans. Pattern Anal. and Machine Intell.*, 31(11):2000–2014.

Reitmaier, T. and Sick, B. (2013). Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS. *Information Sciences*, 230:106–131.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, pages 189–196, Cambridge, MA.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Morgan & Claypool, San Rafael, CA.