IEIE Transactions on Smart Processing and Computing

A Hybrid Selection Method of Helpful Unlabeled Data Applicable for Semi-Supervised Learning Algorithm

Thanh-Binh Le[1] and Sang-Woon Kim[1]

[1] Department of Computer Engineering, Myongji University / 116, Myongji-ro, Cheoin-gu, Yongin-shi, Gyeonggi-do, Korea
{binhle, kimsw}@mju.ac.kr

***** Corresponding Author: Sang-Woon Kim

Received December 20, 2013; Revised February 27, 2013; Accepted May 12, 2014; Published August 31, 2014

** Short Paper*

** Extended from a Conference: Preliminary results of this paper were presented at the ISCE2014, June 22-25, 2014. This present paper has been accepted by the editorial board through the regular reviewing process that confirms the original contribution.*

Abstract**:** This paper presents an empirical study on selecting a small amount of useful unlabeled data to improve the classification accuracy of semi-supervised learning algorithms. In particular, a hybrid method of unifying the simply recycled selection method and the incrementally-reinforced selection method was considered and evaluated empirically. The experimental results, which were obtained from well-known benchmark data sets using semi-supervised support vector machines, demonstrated that the hybrid method works better than the traditional ones in terms of the classification accuracy.

Keywords**:** Pattern recognition and machine learning, Semi-supervised learning (SSL), Simply recycled selection (SRS), Incrementally reinforced selection (IRS), Hybrid selection strategy (HYB).

1. Introduction

The semi-supervised learning (SSL) approach [1] is one way of addressing the insufficiency of labeled data in pattern recognition and machine learning. That is, in SSL, both a limited number of labeled data and a multitude of unlabeled data are used to learn a classification model. The utilization of the unlabeled data, however, is not always helpful for SSL algorithms [2, 3]. When classifier designers have more data for learning, they have more information for classification. On the other hand, having many more data examples also implies an increased likelihood of having more bad information, particularly, when they have the wrong link assumptions between the marginal distribution and the conditional distribution. Therefore, to efficiently use the unlabeled examples in learning the classification model, some examples deemed useful for the learning process can be selected and given the correctly estimated labels.

To address this concern, especially when dealing with semi-supervised support vector machines (S3VMs) [4], two selection strategies, named the simply recycled selection (SRS) and the incrementally reinforced selection (IRS) methods, have recently been considered and compared empirically [5, 6]. In IRS, a small portion of strong examples are selected from the available unlabeled data set in an incremental fashion. In SRS, the amount of the selected examples is fixed over boosting iterations [7]. On the other hand, in IRS, certain kinds of selected data that have been evaluated appropriately but given incorrect-prediction labels or vice versa continue to be used in the next iteration steps; which means that learning leads to poor classification performance. To remedy this problem, a hybrid method, composed of SRS and IRS methods, was considered in this paper (see Figs. 1 and 2). The experimental results, which were obtained using well-known benchmark data sets through semi-supervised support vector machines, demonstrate that the proposed method can compensate for the shortcomings of traditional algorithms.

The remainder of the paper is organized as follows. Section 2 briefly explains the methods for improving S3VMs by utilizing the modified criterion for selecting a small amount of helpful unlabeled data. Section 3 presents the proposed hybrid method. Section 4 describes the experimental setup and presents the results. Finally, Section 5 provides the conclusion.

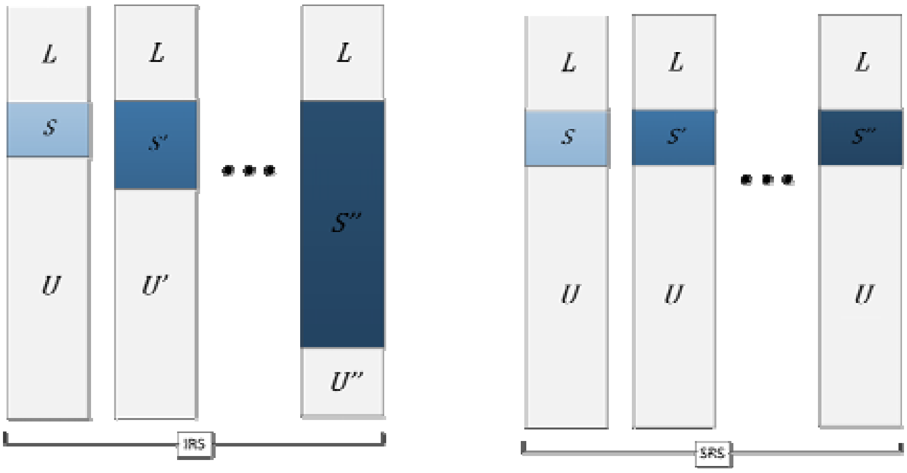
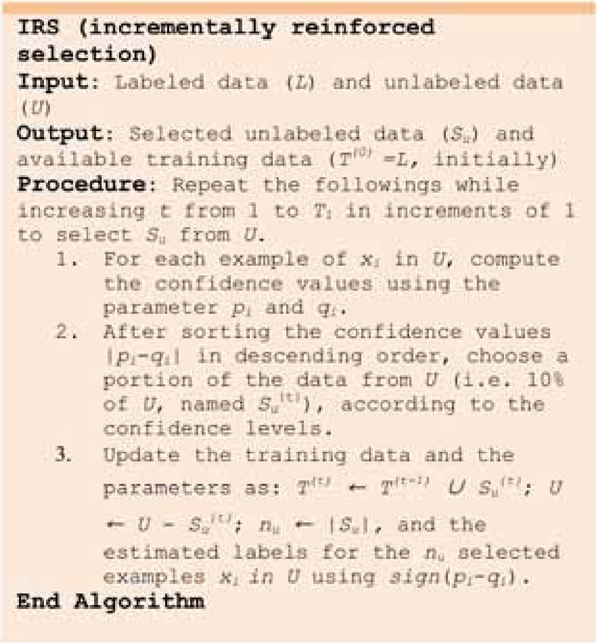


Fig. 1. Plots comparing the two strategies for selection. From the left (a) incrementally reinforced selection (IRS), (b) simply recycled selection (SRS).

2. Related Work

In this section, IRS and SRS, which are two selection strategies to be combined in the present empirical study, are reviewed briefly. Details of the algorithms can be found elsewhere [5, 7].

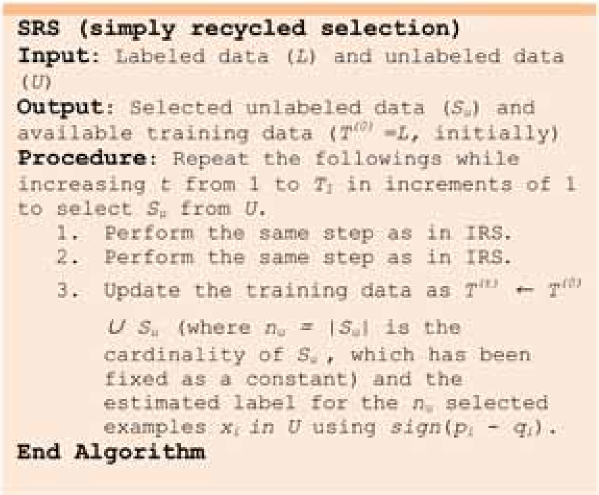
The IRS procedure of selecting a small number of the useful samples from the available unlabeled ones is summarized below. Here, the labeled data (L) and unlabeled data (U) are given as input parameters. After performing the sampling procedure, (newly updated) training data $T^{(t)}$ and (newly updated) unlabeled data S_u are obtained as outputs. The parameters of p, q, and P_E (estimated conditional class probability), however, are calculated internally using a similarity matrix between the pair-wise examples, L and U:



```
IRS (incrementally reinforced
selection)
Input: Labeled data (L) and unlabeled data
(U)
Output: Selected unlabeled data (S_u) and
available training data (T^(0) =L, initially)
Procedure: Repeat the followings while
increasing t from 1 to T_1 in increments of 1
to select S_u from U.
    1.  For each example of x_i in U, compute
        the confidence values using the
        parameter p_i and q_i.
    2.  After sorting the confidence values
        |p_i-q_i| in descending order, choose a
        portion of the data from U (i.e. 10%
        of U, named S_u^(t)), according to the
        confidence levels.
    3.  Update the training data and the
        parameters as: T^(t) ← T^(t-1) ∪ S_u^(t); U
        ← U - S_u^(t); n_u ← |S_u|, and the
        estimated labels for the n_u selected
        examples x_i in U using sign(p_i-q_i).
End Algorithm
```

From the above algorithm, in step 1, 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. Please refer to [5, 7] for further details.

SRS strategy is summarized as follows:



```
SRS (simply recycled selection)
Input: Labeled data (L) and unlabeled data
(U)
Output: Selected unlabeled data (S_u) and
available training data (T^(0) =L, initially)
Procedure: Repeat the followings while
increasing t from 1 to T_1 in increments of 1
to select S_u from U.
    1.  Perform the same step as in IRS.
    2.  Perform the same step as in IRS.
    3.  Update the training data as T^(t) ← T^(0)
        ∪ S_u (where n_u = |S_u| is the
        cardinality of S_u , which has been
        fixed as a constant) and the
        estimated label for the n_u selected
        examples x_i in U using sign(p_i - q_i).
End Algorithm
```

In addition to these descriptions, Fig. 1 presents a graphical comparison of the two strategies. The IRS and SRS strategies are presented from the left to right rows [5].

3. Hybrid Selection Strategy

In this section, two selection strategies, i.e. SRS and IRS, are combined into a hybrid algorithm (shortly HYB). Using the HYB strategy, strongly discriminative examples are first selected from unlabeled data, and together with labeled data, utilized for training a (supervised) classifier or used for retraining the ensemble classifier. In this scenario, a classifier (H) can be trained using a small number unlabeled examples selected from the unlabeled data set, U, and the labeled data set, L. This selecting-and-training process is repeated for a predefined number of iterations or until a termination criterion is met.

Initially, the parameters, such as $\alpha^{(0)}$ (selection rate), $\Delta^{(0)}$ (incremental selection rate), Φ (kernel function), T_1

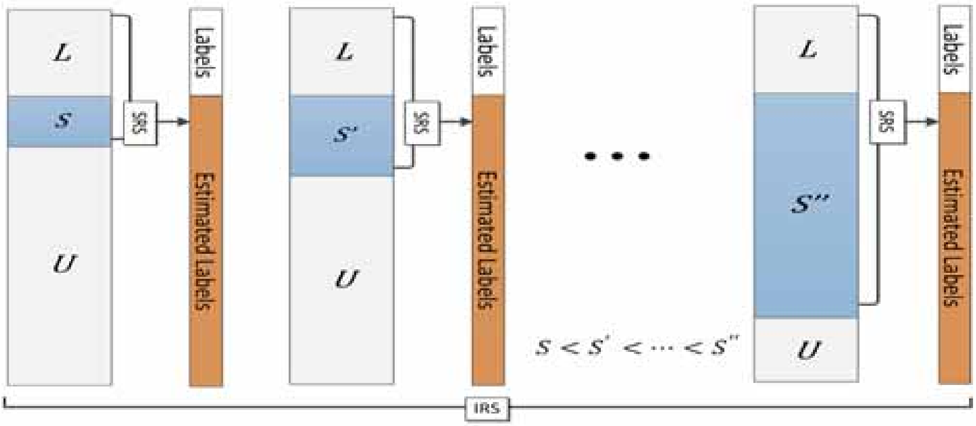
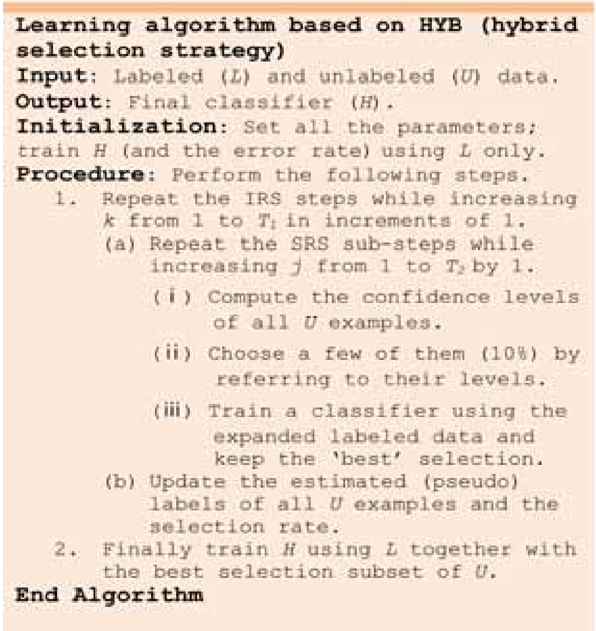


Fig. 2. Learning algorithm based on the proposed hybrid selection strategy, where SRS and IRS are presented in Fig. 1.

and T_2 (predetermined iteration numbers), and $h^{(0)}$ (a base classifier, e.g. SVM), are set. After selecting $S_u^{(0)}$ (a subset) from U to $h^{(0)}$, which has been trained with L only, train a S3VM (H) using $\{L \cup S_u^{(0)}\}$ and calculate the training error rate ($\varepsilon(H)$) using L only. After performing the SRS sub-steps while increasing j from 1 to T_2 in increments of 1, repeat the IRS sub-steps while increasing k from 1 to T_1 in increments of 1. That is, for each $x_i \in U$, first compute the classification confidence levels ($|CL(x_i)|$) using p_i, q_i, $P_E(x_i)$, and $h^{(j-1)}$ (refer to [5, 7] for the details). By referring to the confidence levels of all $x \in U$, choose $d^{(k)}$ (%) top of U as $S_u^{(j)}$. Next, train a new SVM ($h^{(j}$) with the expanded labeled data (i.e., $\{L \cup S_u^{(j)}\}$) Evaluate and find the best classifier ($h^{(j^*)}$) among all the trained classifiers by the iteration, including $h^{(j)}$.

After updating the estimated (pseudo) labels of U examples, $sign(CL(x_i))$, and the selection rate (i.e., $d^{(k+1)} \leftarrow d^{(k)} + \Delta^{(k)}$) for all the j and k variables, train S3VM (H) again using the finally expanded labeled data $\{L \cup S_u^{(j^*)}\}$.

The HYB-based learning algorithm can be summarized systematically as follows:



```
Learning algorithm based on HYB (hybrid
selection strategy)
Input: Labeled (L) and unlabeled (U) data.
Output: Final classifier (H).
Initialization: Set all the parameters;
train H (and the error rate) using L only.
Procedure: Perform the following steps.
   1.  Repeat the IRS steps while increasing
       k from 1 to T1 in increments of 1.
       (a) Repeat the SRS sub-steps while
           increasing j from 1 to T2 by 1.
           (i)   Compute the confidence levels
                 of all U examples.
           (ii)  Choose a few of them (10%) by
                 referring to their levels.
           (iii) Train a classifier using the
                 expanded labeled data and
                 keep the 'best' selection.
       (b) Update the estimated (pseudo)
           labels of all U examples and the
           selection rate.
   2.  Finally train H using L together with
       the best selection subset of U.
End Algorithm
```

Fig. 2 presents a plot explaining a learning algorithm based on the proposed selection strategy, where learning is carried out from the left to right steps while increasing the cardinality of the selected unlabeled subset. First, a subset S, which corresponds to S_u, is selected from U in SRS fashion. After updating the estimated labels, another subset S' is selected from U in SRS fashion, where the cardinality of S' is greater than that of S. These steps are repeated in an IRS manner.

4. Experimental Results

This section reports the run-time characteristics of the algorithm to illustrate the functioning of the newly proposed strategy. The proposed HYB selecting and learning strategy was evaluated using three sets of data: UCI data (four datasets) [8], SSL-Book benchmark data (four datasets) [9], and Practical Image datasets of VOC'07 data (five datasets) [10, 11]. Qualification of the VOC'07 image data sets was verified using well-known PASCAL VOC'07 data. Table 1 lists the characteristics of the experimental data.

The proposed strategy was tested and compared with the conventional strategies reported in the literature. This was achieved by performing the following experiments. First, each dataset was divided into three subsets: the labeled training subset, L, the evaluation test subset, E, and the unlabeled subset, U, at a ratio of 20(%): 20(%): 60(%). The training and evaluation procedures were then repeated *100* times. Finally, the results obtained were averaged. Table 1 presents the classification error rates (mean values and standard deviations) (%). The results shown in the third column were obtained with the HYB (hybrid selection) method, whereas those of the fourth and fifth column were obtained using the IRS (incrementally reinforced selection) and SRS (simply recycled selection), respectively. In the experiment, the cardinality of the selected subset at iteration is n_u=10(%). In addition, all of the S3VM algorithms were implemented using publicly available software [4].

Therefore, for a simple comparison, the number of * markers that each algorithm earned for the thirteen datasets

Table 1. Experimental data characteristics.

Dataset types	Dataset names	# of dimens	# of classes	# of objects
UCI	Heart	13	2	297
	Breast	9	2	683
	Ionosphere	34	2	351
	Diabetes	8	2	768
SSL-Book	Digit1	241	2	1500
	BCI	117	2	400
	USPS	241	2	1500
	g241n	241	2	1500
VOC'07	Aeroplane	4000	2	1131
	Motorbike	4000	2	1139
	Person	4000	2	2044
	Car	4000	2	1395
	Horse	4000	2	1158

Table 2. Numerical comparison of the mean error rates (and standard deviations) (%). For the ease of comparison, the lowest rate in each data was highlighted with a * marker.

Data types	Data names	Error rates (SD)		
		HYB	IRS	SRS
UCI	Heart	9.67 (3.99)	*9.66 (4.20)	14.56 (3.20)
	Breast	*4.32 (1.72)	5.36 (1.65)	5.42 (1.75)
	Ionosphere	8.00 (3.71)	*7.63 (3.23)	10.63 (3.04)
	Diabetes	35.42 (2.01)	*35.10 (1.44)	38.20 (3.40)
SSL-Book	Digit1	*2.53 (0.91)	2.54 (0.85)	6.55 (2.86)
	BCI	*47.48 (4.43)	49.83 (1.92)	49.80 (1.90)
	USPS	*9.14 (2.59)	9.79 (2.25)	12.54 (2.35)
	g241n	*48.36 (2.47)	49.95 (2.16)	49.95 (2.16)
VOC'07	Aeroplane	*6.40 (1.72)	*6.53 (1.25)	8.63 (1.12)
	Motorbike	*9.77 (1.39)	9.89 (1.37)	10.54 (1.05)
	Person	*35.02 (2.14)	*35.02 (2.14)	42.00 (5.40)
	Car	*22.12 (3.11)	22.17 (3.26)	27.22 (3.47)
	Horse	*9.82 (1.59)	9.95 (0.19)	11.05 (1.35)

were counted and compared. Based on this evaluation system, the ranks of the three algorithms were: HYB (10), IRS (5), and SRS (0), where the numbers in (•) denote the number of * markers that the algorithms obtained in the competition.

This result means that in S3VMs, there is no specific approach that yields the best results for all families of applications in terms of the classification accuracy. The best classifier and/or SSL approach for one dataset is not the best for another dataset. On the other hand, the accuracy of HYB is slightly higher than that of the others; or the accuracies of HYB and IRS are similar.

In addition to this simplistic comparison, to demonstrate the significant differences in the error rates between the three approaches used in the experiments, for the classification mean (and standard deviation) rates (%) shown in Table 2, the Wilcoxon signed-rank test [12], which is a non-parametric statistical hypothesis test used to compare two related samples (e.g. error rates of two classifiers), can be conducted. In particular, to compare the classification accuracies of the classifiers designed with different selection strategies, the following steps are performed. First, let $p^{(i)} = p_A^{(i)} - p_B^{(i)}$ be the difference between the performance scores of two classifiers A and B on the i^{th} out of N data sets. Rank $p^{(i)}$, $i = 1, ..., N$ according to their absolute values. Compute the positive rank sum (R^+) and negative rank sum (R^-), as defined in [12]. Find a critical value of Wilcoxon signed-rank test (by referring to Table 8 in [13]) and set $T = \min(R^+, R^-)$. Finally, hypothesis H_0 could be rejected if the observed value of T is smaller than the critical value.

First, for HYB vs. IRS, set the hypothesis as follows:

H_0: HYB is equal to IRS.

H_1: HYB is significantly different with IRS.

Second, for HYB vs. SRS, set the hypothesis as follows:

H_0: HYB is equal to SRS.

H_1: HYB is significantly different with SRS.

Tables 3 and 4 present the values of the $p^{(i)}$ measured for HYB vs. IRS and HYB vs. SRS, respectively.

From Table 3 (and Table 4), the positive rank sum, R , and the negative rank sum, R^-, were measured as follows. First, from Table 3, $R^+ = (2 + 9 + 8) + ½(1) = 19.5$, where pairs with $p^{(i)} < 0$ are excluded and the sample size is reduced. $R^- = (11+3+13+10+12+7+5+4+6) + ½(1) = 71.5$, where pairs with $p^{(i)} > 0$ are excluded; $T = \min(R+, R^-) = 19.5$; $N = 13$. The critical value is 21 at one-tailed ($\alpha = 0.05$). $T <$ Critical value, which means that H_0 is *rejected*, but H_1 is *accepted*; i.e., HYB is significantly different from IRS. Next, from Table 4, $R^+ = 0$; $R^- = 91$, $T = \min(R^+, R^-) = 0$, $N = 13$. The critical value is 21 at one-tailed ($\alpha = 0.05$). $T <$ Critical value, which means that H_0 is also *rejected*, whereas H_1 is *accepted*. Therefore, HYB is significantly different from SRS.

Table 3. Values of $p^{(i)}$ measured with HYB vs. IRS.

Data names	Error rates		$Pp^{(i)}$	Rank | $p^{(i)}$|
	HYB	IRS		
Heart	9.67	9.66	0.01	2
Breast	4.32	5.36	-1.04	11
Ionosphere	8	7.63	0.37	9
Diabetes	35.42	35.1	0.32	8
Digit1	2.53	2.54	-0.01	3
BCI	47.48	49.83	-2.35	13
USPS	9.14	9.79	-0.65	10
g241n	48.36	49.95	-1.59	12
Aeroplane	6.4	6.53	-0.13	7
Motorbike	9.77	9.89	-0.12	5
Person	35.02	35.02	0	1
Car	22.12	22.17	-0.05	4
Horse	9.82	9.95	-0.13	6

Table 4. Values of $p^{(i)}$ measured with HYB vs. SRS.

Data names	Error rates		$Pp^{(i)}$	Rank | $p^{(i)}$|
	HYB	SRS		
Heart	9.67	14.56	-4.89	9
Breast	4.32	5.42	-1.10	5
Ionosphere	8	10.63	-2.63	11
Diabetes	35.42	38.2	-2.78	10
Digit1	2.53	6.55	-4.02	12
BCI	47.48	49.8	-2.32	7
USPS	9.14	12.54	-3.40	6
g241n	48.36	49.95	-1.59	3
Aeroplane	6.4	8.63	-2.23	4
Motorbike	9.77	10.54	-0.77	2
Person	35.02	42	-6.98	13
Car	22.12	27.22	-5.10	8
Horse	9.82	11.05	-1.23	1

5. Conclusion

This paper reported the results of an empirical study on evaluating the hybrid method (HYB) of the conventional IRS and SRS selection strategies when dealing with semi-supervised support vector machines (S3VMs). Three selection and learning strategies, which utilize the HYB, IRS and SRS strategies, respectively, were evaluated and compared using well-known public domain datasets. The experimental results obtained demonstrate that the classification accuracy of the S3VMs was improved marginally by employing the HYB strategy. In particular, the results of the Wilcoxon signed-rank tests for the classification error rates obtained with the experimental datasets showed that HYB is significantly different from IRS and SRS. Although S3VM-HYB can be improved in terms of the classification accuracy, the experiments performed were limited. Therefore, further studies with various data sets and selection strategies will be needed.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science, ICT and Future Planning) (NRF-2012R1A1A2041661). The authors are grateful to this support.

References

[1] X. Zhu, A. B. Goldberg, Introduction to Semi-Supervised Learning, Morgan & Claypool, San Rafael, CA, 2009 doi:10.2200/S00196ED1V01Y200906AIM006

[2] F. G. Cozman, I. Cohen, M. C. Cirelo, "Semi-supervised learning of mixture models," in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003. Available at Article (CrossRef Link).

[3] D. Elworthy, “Does Baum-Welch re-estimation help taggers?” in Proceedings of the fourth conference on Applied natural language processing (ANLC'94) , pp. 53-58, 1994. Doi:10.3115/974358.974371

[4] C. -C. Chang and C. -J. Lin, “LIBSVM: a library for support vector machines,” Journal ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, pp. 1-27, 2011. doi:10.1145/1961189.1961199

[5] T. -B. Le and S. -W. Kim, “On incrementally using a small portion of strong unlabeled data for semi-supervised learning algorithms,” Pattern Recognition Letters, vol. 41, pp. 53-64, May 2014. doi:10.1016/j.patrec.2013.08.026

[6] T. -B. Le and S. -W. Kim, “Simply recycled selection and incrementally reinforced selection methods applicable for semi-supervised learning algorithms,” in *Proc. of the 2014 Int'l Conf. on Electronics, Information and Communication (ICEIC 2014)*, pp. 15-18, Jan. 2014.

[7] P. K. Mallapragada et al., “SemiBoost: Boosting for semi-supervised learning,” *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 312, no. 11, pp. 2000-2014, Nov. 2009. doi: 10.1109/TPAMI.2008.235.

[8] A. Asuncion and D. J. Newman, “UCI Machine Learning Repository,” *Technical Report, University of California, School of Information and Computer Science*, Irvine, CA, 2007. Available at Article (CrossRef Link)

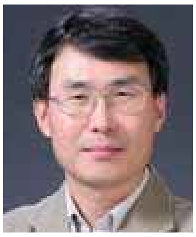
[9] O. Chapelle et al., “Semi-Supervised Learning,” *The MIT Press*, MA, 2006. Available at Article (CrossRef Link).

[10] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman, “The PASCAL Visual Object Classes (VOC) Challenge,” Int J Comput Vis (2010) vol. 88. pp. 303–338, 2010. Available at Article (CrossRef Link) doi:10.1007/s11263-009-0275-4.

[11] A. Vedaldi et al., “Image Classification Practical 2011,” *The MIT Press*, MA, 2006, Available at Article (CrossRef Link)

[12] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83, Dec 1945. Available at Article (CrossRef Link)

Thanh-Binh Le received his B.S. degree in Computer Science and Engineering from The University of Pedagogy - HCMC, Vietnam, in 2009, and the ME degree from Myongji University, Yongin, Korea in 2012, in Computer Engineering. Currently, he is a PhD student of Department of Computer Engineering, Myongji University, Korea. His research interests include Pattern Recognition and Machine Learning.



Sang-Woon Kim received the BE degree from Hankook Aviation University, Gyeonggi, Korea in 1978, and the ME and the PhD degrees from Yonsei University, Seoul, Korea in 1980 and 1988, respectively, both in Electronic Engineering. In 1989, he joined the Department of Computer Science and Engineering, Myongji University, Korea and is currently a Full Professor there. His research interests include Statistical Pattern Recognition, Machine Learning, and Avatar Communications in Virtual Worlds. He is the author or coauthor of 40 regular papers and 13 books. He is a Senior Member of the IEEE and a member of the IEIE.