Unsupervised Ensemble Learning for Class Imbalance Problems

Zihan Liu, Dongrui Wu, *Senior Member, IEEE*

Key Laboratory of the Ministry of Education for Image Processing and Intelligent Control
School of Automation, Huazhong University of Science and Technology, Wuhan, China
E-mail: zhliu95@hust.edu.cn, drwu@hust.edu.cn

***Abstract*—Ensemble learning, which aggregates multiple base (weak) learners to obtain a strong learner, is an effective approach for improving the generalization performance of a machine learning model. Several completely unsupervised ensemble learning approaches have been proposed in the literature for binary classification. However, most of them only considered the case that the two classes are balanced, and hence their performances deteriorate when there is significant class imbalance, which often happens in practice. This paper proposes a spectral meta-learner for class imbalance (SMLCI) approach to explicitly consider the class imbalance. Experiments on 12 UCI datasets from various domains verified that SMLCI significantly outperformed the individual base classifiers, and also five existing unsupervised ensemble learning approaches, when the balanced classification accuracy is used as the performance measure.**

***Keywords*—Ensemble learning; unsupervised learning; class imbalance; binary classification**

I. Introduction

Ensemble learning [7], [22] is a popular machine learning approach for aggregating multiple base (weak) learners to obtain a strong learner. Compared with each individual base learner, the strong learner usually has a smaller variance and better generalization performance.

In practice the strong learner may be applied to data that the base learners have not seen before. For example, in a brain-computer interface application studied in [21], several base learners were built from labeled EEG trials from some existing subjects, and then the aggregated strong learner was applied to unlabeled EEG trials from a new subject. This represents a completely unsupervised ensemble learning problem, as none of the EEG trials from the new subject was labeled. On the other hand, if there are some labeled samples from the new subject available to refine the strong learner, then it becomes a supervised or semi-supervised ensemble learning problem. Generally unsupervised ensemble learning is more challenging than supervised or semi-supervised ensemble learning, and this paper considers the former.

According to the characteristics of the base learners, ensemble learning approaches can be categorized into two groups [22]:

1) *Homogeneous ensemble learning*, in which all base learners are of the same type, e.g., random forest for classification and regression [12], where all base learners are decision trees.
2) *Heterogeneous ensemble learning*, in which the base learners are of different types, e.g., the base learners could be a mixture of decision trees [17], neural networks [18], support vector machines [19], etc..

This paper focuses on heterogeneous ensemble learning for binary classification problems. However, our approach can also be applied to homogeneous ensemble learning problems.

According to whether the base learners are dependent on each other or not, ensemble learning approaches can also be categorized into two groups [22]:

1) *Sequential approaches*, where the construction of a base learner depends on the performance of its pervious base learners, e.g., adaBoost [6].
2) *Parallel approaches*, where the base learners are constructed in parallel and independent of each other, e.g., bagging [11].

This paper focuses on the latter. In summary, this paper studies unsupervised heterogeneous ensemble learning for binary classification, where the base learners are constructed in parallel.

Majority vote (MV), which assigns a sample to the class that the most base classifiers agree on, is perhaps the most widely-used ensemble learning approach for classification. In its simplest form, all base classifiers have equal weights. However, in practice the base classifiers usually have different accuracies, and it is more intuitive to use weighted voting to emphasize better classifiers. The challenge is how to optimally design such weights, especially in unsupervised ensemble learning, where only unlabeled samples are available.

Recently a few more sophisticated unsupervised heterogeneous ensemble learning approaches for binary classification have been proposed. Parisi et al. [15] proposed a spectral meta-learner (SML) approach to estimate the accuracies of the base classifiers from their population covariance matrix, and then used them in a maximum likelihood estimator (MLE) to aggregate these base classifiers. Researchers from the same group then proposed several improvements [9], [10] to the SML. Wu et al. [20] proposed an agreement rate initialized maximum likelihood estimator (ARIMLE), which used a simplified agreement rate (AR) approach [16] to estimate the classification accuracy of each base classifier from the unlabeled samples, initialized an MLE from these accuracies, and then iterated the MLE to obtain better performance.

978-1-7281-1312-8/18/$31.00 ©2018 IEEE

The above approaches have demonstrated promising performances in various applications. However, most of them assume that the two classes are balanced, i.e., the positive class has roughly the same number of samples as the negative class. In practice usually it is difficult to know *a priori* if this assumption is true. In fact, most real-world applications have more or less class imbalance [8], [14]. The performances of the above approaches may deteriorate in such cases.

This paper proposes a spectral meta-learner for class imbalance (SMLCI) approach to explicitly cope with class imbalance. Experiments on 12 UCI datasets from various domains demonstrated the effectiveness of the SMLCI.

The remainder of this paper is organized as follows: Section II introduces several representative existing unsupervised heterogeneous ensemble learning approaches. Section III proposes our SMLCI. Section IV compares the performance of SMLCI with eight individual base classifiers, and also five existing unsupervised heterogeneous ensemble learning approaches. Finally, Section V draws conclusions.

II. Existing Unsupervised Heterogeneous Ensemble Learning Approaches

This section introduces five representative existing unsupervised heterogeneous ensemble learning approaches, which will be compared with our proposed SMLCI in Section IV.

A. Problem Setup

We consider binary classification. Let $\mathcal{X}$ be the input space and $\mathcal{Y} \in \{-1, 1\}$ be the output space. A labeled sample $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is a random vector with joint probability density $p(\mathbf{x}, y)$. Let $\{f_i\}_{i=1}^m$ be m classifiers operating on $\mathcal{X}$, and assume there are n unlabeled samples, $\{\mathbf{x}_j\}_{j=1}^n$, with unknown true labels $\{y_j\}_{j=1}^n$. Define the classification sensitivity of f_i as

$$\psi_i = \mathrm{P}(f_i(X) = 1 | Y = 1) \tag{1}$$

and its specificity as

$$\eta_i = \mathrm{P}(f_i(X) = -1 | Y = -1) \tag{2}$$

Then, the balanced classification accuracy (BCA) of f_i is computed as [9], [10], [15], [20]:

$$\pi_i = \frac{1}{2}(\psi_i + \eta_i). \tag{3}$$

The BCA is used as the performance measure in this paper, as it emphasizes the classification accuracies of both classes equally, which is more appropriate than the overall classification accuracy (the number of correctly classified samples divided by the total number of samples) for class imbalance problems.

B. Majority Vote (MV)

MV is perhaps the simplest and also most widely-used ensemble learning approach for classification. It assigns a sample to the class that the most base classifiers agree on. Mathematically,

$$\hat{y}_j = \mathrm{sign}\left[\sum_{i=1}^m f_i(\mathbf{x}_j)\right], \quad j = 1, ..., n. \tag{4}$$

C. Spectral Meta-Learner (SML)

SML [15] is a linear approximation to the MLE, which has demonstrated promising performance in many applications [15], [21]. It uses two important assumptions: 1) The n unlabeled samples $\{\mathbf{x}_j\}_{j=1}^n$ are independent and identically distributed realizations from $p_X(\mathbf{x})$; and, 2) The m base binary classifiers $\{f_i\}_{i=1}^m$ are conditionally independent.

As shown in [15], the MLE is a linear combination of the binary labels from the base classifiers:

$$\hat{y}_j = \mathrm{sign}\left[\sum_{i=1}^m (f_i(\mathbf{x}_j) \ln \alpha_i + \ln \beta_i)\right] \tag{5}$$

where the weights and the bias depend on the unknown specificities and sensitivities of the m base classifiers, i.e.,

$$\alpha_i = \frac{\psi_i \eta_i}{(1 - \psi_i)(1 - \eta_i)} \tag{6}$$

$$\beta_i = \frac{\psi_i (1 - \psi_i)}{\eta_i (1 - \eta_i)} \tag{7}$$

A Taylor expansion of the unknown coefficients α_i and β_i in (6) and (7) around $(\psi_i, \eta_i) = (1/2, 1/2)$ gives:

$$\alpha_i \approx 1 + 4(\psi_i + \eta_i - 1) = 1 + 4(2\pi_i - 1) \tag{8}$$

$$\beta_i \approx 1 \tag{9}$$

Hence, (5) can be rewritten as:

$$\hat{y}_j \approx \mathrm{sign}\left[\sum_{i=1}^m f_i(\mathbf{x}_j)(2\pi_i - 1)\right] \tag{10}$$

Denote $\hat{\mathbf{y}}_i = [f_i(\mathbf{x}_1), ..., f_i(\mathbf{x}_n)]^T$, $i = 1, ..., m$, and $Y - [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_m]$. Then, the covariance matrix of the m binary classifiers is computed as:

$$Q = Y^T Y \tag{11}$$

Let the leading eigenvector be $\mathbf{v} = [v_1, ..., v_m]$. It has been shown [15] that $v_i \propto 2\pi_i - 1$. Then, (10) can be re-formulated as:

$$\hat{y}_j \approx \mathrm{sign}\left[\sum_{i=1}^m f_i(\mathbf{x}_j) \cdot v_i\right], \quad j = 1, ..., n \tag{12}$$

The complete SML algorithm is shown in Algorithm 1.

D. Improved Spectral Meta-Learner (i-SML)

The derivation of SML uses an assumption that $(\psi_i, \eta_i) = (1/2, 1/2)$. Thus, it is sub-optimal when few classifiers are significantly more accurate than all others. So, Jaffe et al. [10] proposed an improved spectral meta-learner (i-SML) by first estimating the class imbalance $b = P(Y = 1) - P(Y = -1)$ of the labels and then using it to directly estimate the sensitivity and specificity of each base classifier.

Algorithm 1: The SML algorithm [15].

Input: n unlabeled samples, $\{\mathbf{x}_j\}_{j=1}^n$;
m base binary classifiers, $\{f_i\}_{i=1}^m$.
Output: The n estimated class labels, $\hat{y}_j$.
Apply each of $\{f_i\}_{i=1}^m$ to the n unlabeled samples and obtain the predictions $f_i(\mathbf{x}_j)$, $i = 1, ..., m$, $j = 1, ..., n$;
Compute the covariance matrix Q in (11);
Compute the first leading eigenvector, $\mathbf{v}$, of Q;
Return $\hat{y}_j$ computed by (12).

The class imbalance b can be estimated by a 3D covariance tensor approach or a restricted-likelihood approach [10]. Once b is obtained, ψ_i and η_i can be computed as:

$$\psi_i = \frac{1}{2}\left(1 + \mu_i + v_i\sqrt{\frac{1-b}{1+b}}\right) \tag{13}$$

$$\eta_i = \frac{1}{2}\left(1 - \mu_i + v_i\sqrt{\frac{1+b}{1-b}}\right) \tag{14}$$

where $\mathbf{v} = [v_1, ..., v_m]$ is the leading eigenvector of Q in (11), μ_i is the sample mean of the ith base classifier, i.e.,

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} f_i(\mathbf{x}_j) \tag{15}$$

(13) and (14) are then substituted into (6) and (7) to compute α_i and β_i, which are next substituted into (5) to compute the final class labels.

E. Latent Spectral Meta-Learner (L-SML)

Latent SML (L-SML) [9] is an unsupervised model that allows for dependencies between the base classifiers. It assumes the m base classifiers can be partitioned into several groups according to a latent variable: the classifiers in the same group can be correlated, but these from different groups are conditionally independent. L-SML has two steps [9]:

1) Estimate the model parameters. It proceeds in two stages: (i) Estimate the sensitivity and specificity of the base classifiers given the latent variables α_k ($k = 1, ..., K$, $K < m$); and (ii) estimate the probabilities associated with the latent variables $\mathrm{P}(\alpha_k = 1|Y = 1)$ and $\mathrm{P}(\alpha_k = -1|Y = -1)$.
2) Label predictions. The label of $\mathbf{x}_j$ can be estimated by the MLE:

$$\hat{y}_j = \arg\max_{y=\pm 1} \mathrm{P}(f_1(\mathbf{x}_j), \ldots, f_m(\mathbf{x}_j)|y) \tag{16}$$

F. ARIMLE

ARIMLE [20] is also an unsupervised ensemble learning approach to construct a strong learner from several base binary classifiers. It first uses the AR method [16] to estimate the error rate of each base classifier from the unlabeled samples, then computes the accuracies from the error rates and employs them to initialize an MLE, and computes the final prediction by using an EM algorithm to refine the MLE.

The error rate of a binary classifier is defined as:

$$e_i = P(f_i(X) \neq Y), \quad i = 1, \ldots, m \tag{17}$$

Define the AR of two classifiers f_{i_1} and f_{i_2} ($i_1 \neq i_2$) as the probability that they give identical outputs [16], i.e.,

$$a_{i_1,i_2} = P(f_{i_1}(X) = f_{i_2}(X)) = 1 - e_{i_1} - e_{i_2} + 2e_{i_1,i_2} \tag{18}$$

where e_{i_1,i_2} is the joint error rate of f_{i_1} and f_{i_2}. Under the assumption that f_{i_1} and f_{i_2} are independent, we have $e_{i_1,i_2} = e_{i_1} \cdot e_{i_2}$, and hence (18) can be re-expressed as:

$$a_{i_1,i_2} = 1 - e_{i_1} - e_{i_2} + 2e_{i_1} \cdot e_{i_2} \tag{19}$$

Therefore it's feasible to find the m error rates for the m binary classifiers by computing a_{i_1,i_2} for all $\frac{1}{2}m(m-1)$ possible combinations of (i_1, i_2), $i_1 = 1, \ldots, m$, $i_2 = 1, \ldots, m$, and $i_1 \neq i_2$, which can be easily solved by a constrained optimization routine [20].

Once getting the error rates $\{e_1\}_{i=1}^m$, assuming the classification accuracies of positive and negative classes are similar, we have:

$$\pi_i \approx 1 - e_i, \quad i = 1, \ldots, m \tag{20}$$

Substituting (20) into (10), a good initialization of the labels $\{\hat{y}_j\}_{j=1}^m$ can be obtained, i.e.,

$$\hat{y}_j = \mathrm{sign}\left[\frac{\sum_{i=1}^m (2\pi_i - 1) f_i(\mathbf{x}_j)}{\sum_{i=1}^m (2\pi_i - 1)}\right], \quad j = 1, ..., n \tag{21}$$

Then an EM algorithm can be used to refine the MLE. After several iterations, we can get final predicted class labels for the unlabeled samples. It usually gets better performance than initial predictions.

The complete ARIMLE algorithm is shown in Algorithm 2.

Algorithm 2: The ARIMLE algorithm [20].

Input: n unlabeled samples, $\{\mathbf{x}_j\}_{j=1}^n$;
m base binary classifiers, $\{f_i\}_{i=1}^m$.
Output: The maximum likelihood estimates $\{\hat{y}_j\}_{j=1}^n$.
for $i_1 = 1, ..., m-1$ **do**
  for $i_2 = i_1 + 1, ..., m$ **do**
    Compute a_{i_1,i_2} in (19);
  end
  Solve for $\{e_i\}_{i=1}^m$ in (19) using constrained optimization;
  Compute $\{\pi_i\}_{i=1}^m$ using (20);
end
Initialize $\{\hat{y}_j\}_{j=1}^n$ using (21);
while *stopping criterion not met* **do**
  Compute $\{\psi_i\}_{i=1}^m$ in (1) and $\{\eta_i\}_{i=1}^m$ in (2), by treating $\{\hat{y}_j\}_{j=1}^n$ as the true labels;
  Compute $\{\alpha_i\}_{i=1}^m$ in (6) and $\{\beta_i\}_{i=1}^m$ in (7);
  Update $\{\hat{y}_j\}_{j=1}^n$ using (12);
end
Return The latest $\{\hat{y}_j\}_{j=1}^n$.

III. Spectral Meta-Learner for Class Imbalance (SMLCI)

All unsupervised ensemble learning approaches introduced in the previous section, except i-SML, do not explicitly consider class imbalance. However, class imbalance is pervasive in real-world applications, and hence it should not be ignored.

The challenge for handling class imbalance in unsupervised ensemble learning is that the true class imbalance level is unknown, and it cannot be estimated from labeled training samples either, as there are no labeled training samples at all. Thus, classical supervised ensemble learning approaches for handling class imbalance, such as SMOTEBoost [2], EasyEnsemble [13] and BalanceCascade [13], cannot be applied.

This section proposes a simple spectral meta-learner for class imbalance (SMLCI) approach, which uses cost sensitive learning [5] to handle class imbalance. Let r be the (unknown) class imbalance ratio (the number of positive samples divided by the number of negative samples), and $c(i,j)$ be the cost of classifying a sample belonging to Class j to Class i. Then, using the BCA as a performance measure is equivalent to setting $c(1,1) = c(-1,1) = 0$, $c(-1,1) = 1$, and $c(1,-1) = r$. Cost sensitive learning [5] suggests that the optimal decision threshold in this case should be $\frac{r-1}{r+1}$, i.e., (12) should be modified to:

$$\hat{y}'_j = \text{sign}\left[\sum_{i=1}^{m} f_i(\mathbf{x}_j)\cdot v_i - \frac{r-1}{r+1}\right], \quad j = 1,...,n \tag{22}$$

The threshold $\frac{r-1}{r+1}$ becomes 0 when $r = 1$, i.e., when the two classes are balanced, or $c(-1,1) = c(1,-1) = 1$. This is the case in (12). In other words, (12) is optimal when the raw classification accuracy is used as the performance measure, whereas (22) is optimal when the BCA is used as the performance measure.

Unfortunately in unsupervised learning there is no way to know the true class imbalance ratio r. So, we first use SML to estimate the pseudo-labels of the n unlabeled samples, and then compute the estimated class imbalance ratio r from these pseudo labels, i.e.,

$$r = \frac{\sum_{j=1}^{n} I(\hat{y}_j = 1)}{\sum_{j=1}^{n} I(\hat{y}_j = -1)} \tag{23}$$

where $\hat{y}_j$ is computed from (12), and I is an indicator function, i.e.,

$$I(x) = \begin{cases} 1, & x = 1 \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

Then, we substitute r into (22) to update the estimates.

In summary, the pseudo-code of SMLCI is given in Algorithm 3.

IV. Experiments

This section compares the performance of SMLCI with eight base binary classifiers, and also the five existing unsupervised heterogeneous ensemble learning approaches introduced in Section II.

Algorithm 3: The SMLCI approach.

Input: n unlabeled samples, $\{\mathbf{x}_j\}_{j=1}^n$;
m base binary classifiers, $\{f_i\}_{i=1}^m$.
Output: The n estimated class labels, $\hat{y}'_j$.
Apply each of $\{f_i\}_{i=1}^m$ to the n unlabeled samples and obtain the predictions $f_i(\mathbf{x}_j)$, $i = 1,...,m$, $j = 1,...,n$;
Compute the covariance matrix Q in (11);
Compute the first leading eigenvector, $\mathbf{v}$, of Q;
Compute $\hat{y}_j$ in (12);
Compute the class imbalance ratio r in (23);
Return $\hat{y}'_j$ in (22).

A. Experiment Setup

Twelve datasets from the UCI machine learning repository[1], shown in Table I, were used in our experiments. Eight heterogeneous binary classifiers, including logistic regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Multi-layer Perceptron (MLP), and adaBoost (ADA), were trained on 50% randomly selected samples from each dataset. The remaining 50% data were then used for testing. BCA was used as the performance measure. For each dataset we repeated the process 100 times to obtain statistically meaningful results.

TABLE I
Summary of the 12 datasets from the UCI machine learning repository.

Dataset	# Positive Samples	# Negative Samples	Pos/Neg Ratio	# Attributes
Biodeg	356	699	0.51	41
BreastCancer	212	357	0.59	30
Clave	407	4300	0.09	16
ClimateModel	494	46	10.74	18
ILPD	416	167	2.49	10
Ionosphere	225	126	1.79	34
Magic	6688	12332	0.54	10
Musk2	1017	5581	0.18	166
ParkinsonSpeech	520	520	1.00	27
PimaIndiansDiabetes (PID)	268	500	0.54	8
Spambase	1813	2788	0.65	57
Waveform21	1647	3353	0.49	21

B. Performance Comparison with the Base Classifiers

First, we investigated if SMLCI can outperform the eight individual base classifiers. The boxplots of the 100 BCAs on each dataset are shown in Fig. 1, and the average BCAs are shown in Table II. For each dataset the best BCA is marked in bold. Observe that:

1) The individual base classifiers had diverse performances across the 12 datasets, and it is difficult to identify a single base classifier that is always better than others. Hence, using a single base classifier in a specific application may be risky.

[1] https://archive.ics.uci.edu/ml/datasets.html

2) Despite the large variance among the eight base classifiers, ensemble learning by SMLCI almost always outperformed the best base classifier, or achieved comparable performance as the best base classifier. More specifically, Table II shows that SMLCI outperformed all eight base classifiers on six out of the 12 datasets, outperformed seven base classifiers on another five out of the 12 datasets, and outperformed six base classifiers on the one remaining dataset. On average SMLCI outperformed all eight base classifiers.

These results suggested that SMLCI, as an ensemble learning approach, can indeed improve upon the individual base classifiers.

TABLE II
AVERAGE BCAS OF THE EIGHT BASE CLASSIFIERS AND SMLCI ON THE 12 UCI DATASETS.

Dataset	LR	LDA	kNN	SVM	NB	RF	MLP	ADA	SMLCI
Biodeg	.8399	.8318	.8193	.8463	.7076	.8137	.8145	.7804	**.8556**
BreastCancer	.9335	.9390	.7389	.9386	.8640	.9393	.9180	.9153	**.9543**
Clave	.6250	.5289	.6756	.8531	.7323	.7695	.7637	.7435	**.8869**
ClimateModel	.8186	.7190	.5690	.7844	.6071	.5235	.7044	.5908	**.8488**
ILPD	.5747	.5299	.5375	.5249	**.6651**	.5855	.5316	.5940	.6466
Ionosphere	.8170	.8037	.6786	.9059	.9018	**.9388**	.7972	.7792	.9377
Magic	.7452	.7322	.7867	.8157	.6853	**.8357**	.8227	.7790	.8281
Musk2	.8879	.8511	.9115	**.9677**	.8797	.7946	.8993	.8167	.9575
ParkinsonSpeech	**.9996**	.8811	.7924	.9910	.7720	.9404	.9991	.9339	.9846
PID	.7236	.7217	.6713	.7100	.6943	.7124	.7047	.6917	**.7368**
Spambase	.9165	.8700	.8752	.9262	.7891	**.9317**	.9089	.8767	.9313
Waveform21	.8726	.8924	.8074	.8800	.8792	.8919	.8932	.8825	**.9089**
Average	.8128	.7751	.7386	.8453	.7648	.8064	.8131	.7820	**.8731**

We also performed non-parametric multiple comparison tests on the BCAs using Dunn's procedure for multiple comparisons [3], [4], with a p-value correction using the False Discovery Rate method [1]. The corrected p-values are shown in Table III, where the statistically significant ones are marked in bold. Although the eight base classifiers had different performances, the performance improvement of SMLCI over each base classifier was always statistically significant.

TABLE III
p-VALUES OF NON-PARAMETRIC MULTIPLE COMPARISONS OF SMLCI AND THE EIGHT BASE CLASSIFIERS.

	LR	LDA	kNN	SVM	NB	RF	MLP	ADA
LDA	**.0000**							
kNN	**.0000**	**.0000**						
SVM	**.0000**	**.0000**	**.0000**					
NB	**.0000**	**.0000**	**.0001**	**.0000**				
RF	.3191	**.0000**	**.0000**	**.0000**	**.0000**			
MLP	.3333	**.0000**	**.0000**	**.0000**	**.0000**	.4745		
ADA	**0.0000**	.2284	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	
SMLCI	**0.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**

In summary, our experiments confirmed that SMLCI can achieve significantly better BCA than the individual base classifiers.

C. Performance Comparison with Other Unsupervised Ensemble Learning Approaches

Next we compare the performance of SMLCI with the five unsupervised ensemble learning approaches introduced in Section II. Boxplots of the BCAs of the six ensemble learning approaches on the 12 UCI datasets are shown in Fig. 2, and the average BCAs are summarized in Table IV. For each dataset the best BCA is marked in bold. Observe that SMLCI achieved the best average BCA on 11 out of the 12 datasets, and the second best average BCA on the remaining dataset.

TABLE IV
AVERAGE BCAS OF THE SIX UNSUPERVISED ENSEMBLE LEARNING APPROACHES ON THE 12 UCI DATASETS.

Dataset	MV	SML	i-SML	L-SML	ARIMLE	SMLCI
Biodeg	.8468	.8486	.8512	.8494	.8535	**.8556**
BreastCancer	.9398	.9518	.9519	.9491	.9523	**.9543**
Clave	.7003	.7859	.8592	.8577	.8644	**.8869**
ClimateModel	.6978	.7648	.8104	.7946	.7977	**.8488**
ILPD	.5689	.5789	.6149	.6042	.5889	**.6466**
Ionosphere	.9139	.9165	.9327	.9173	.9232	**.9377**
Magic	.7922	.8139	.8193	.8154	.8258	**.8281**
Musk2	.8836	.9102	.9259	.9199	.9371	**.9575**
ParkinsonSpeech	.9673	.9838	.9844	.9628	.9653	**.9846**
PID	.7132	.7211	.7267	.7244	.7281	**.7368**
Spambase	.9245	.9270	.9292	.9286	.9303	**.9313**
Waveform21	.8994	.9062	.9101	.9068	**.9105**	.9089
Average	.8206	.8424	.8597	.8525	.8564	**.8731**

We also performed non-parametric multiple comparison tests on the BCAs using Dunn's procedure for multiple comparisons, with a p-value correction using the False Discovery Rate method. The corrected p-values are shown in Table V, where the statistically significant ones are marked in bold. Observe that:

1) All five more sophisticated ensemble learning approaches (SML, i-SML, L-SML, ARIMLE, and SMLCI) significantly outperformed MV.
2) All four variants of SML (i-SML, L-SML, ARIMLE, and SMLCI) significantly outperformed SML.
3) SMLCI significantly outperformed all other five ensemble learning approaches.

In summary, our experiments further confirmed that SMLCI can achieve significantly better BCA than five existing unsupervised ensemble learning approaches.

TABLE V
p-VALUES OF NON-PARAMETRIC MULTIPLE COMPARISONS OF THE SIX ENSEMBLE LEARNING APPROACHES.

	MV	SML	i-SML	L-SML	ARIMLE
SML	**.0001**				
i-SML	**.0000**	**.0000**			
L-SML	**.0000**	**.0038**	.1174		
ARIMLE	**.0000**	**.0000**	.4746	.1134	
SMLCI	**.0000**	**.0000**	**.0003**	**.0000**	**.0004**

V. CONCLUSIONS

Ensemble learning is an effective approach for improving the generalization performance of a machine learning model. Several completely unsupervised ensemble learning approaches have been proposed in the literature for binary classification. However, most of them can only achieve good performance when the two classes are balanced, i.e., the positive class has roughly the same number of samples as the negative class.

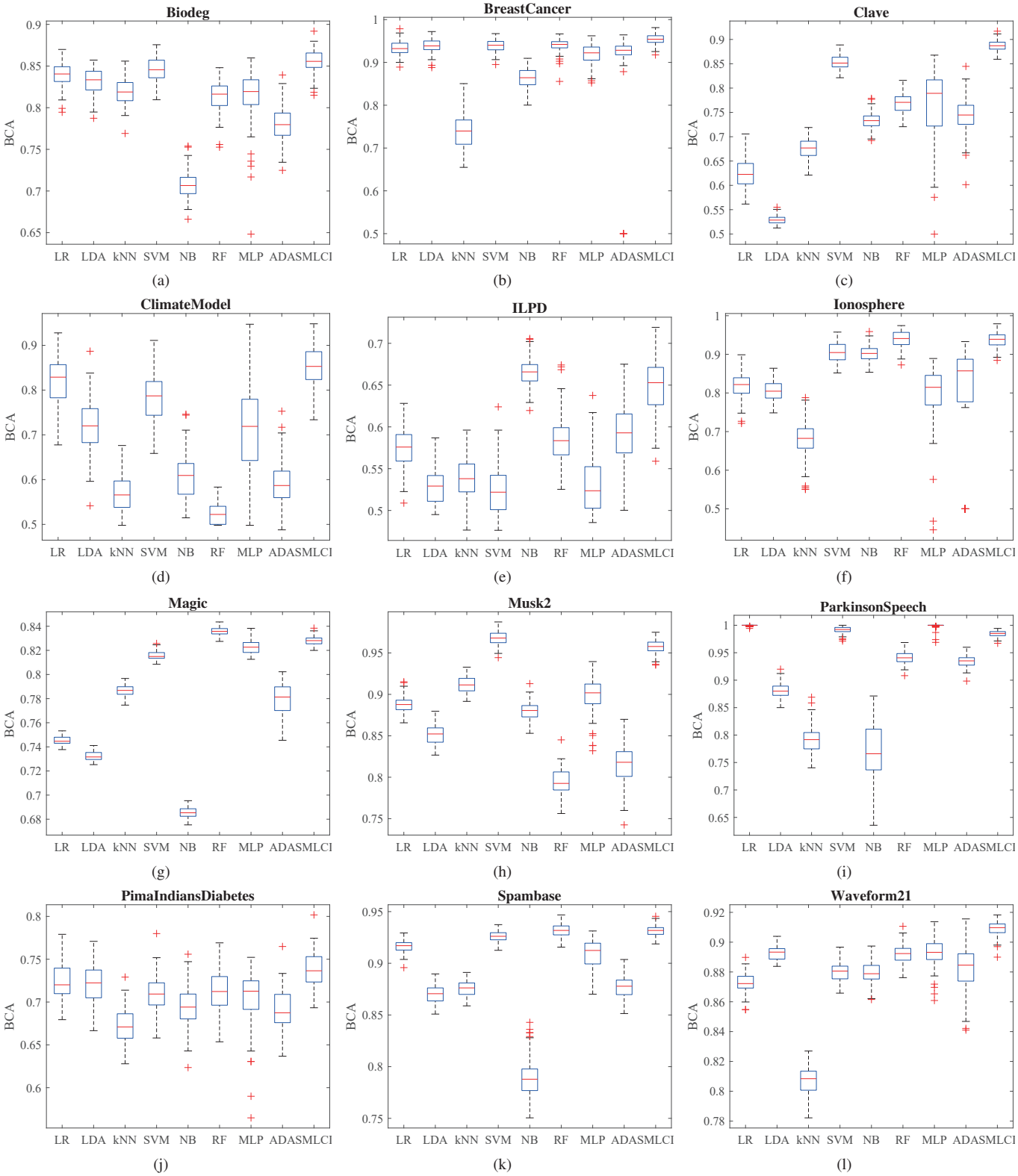

Fig. 1. Performances of the eight base binary classifiers and SMLCI on the 12 UCI datasets.

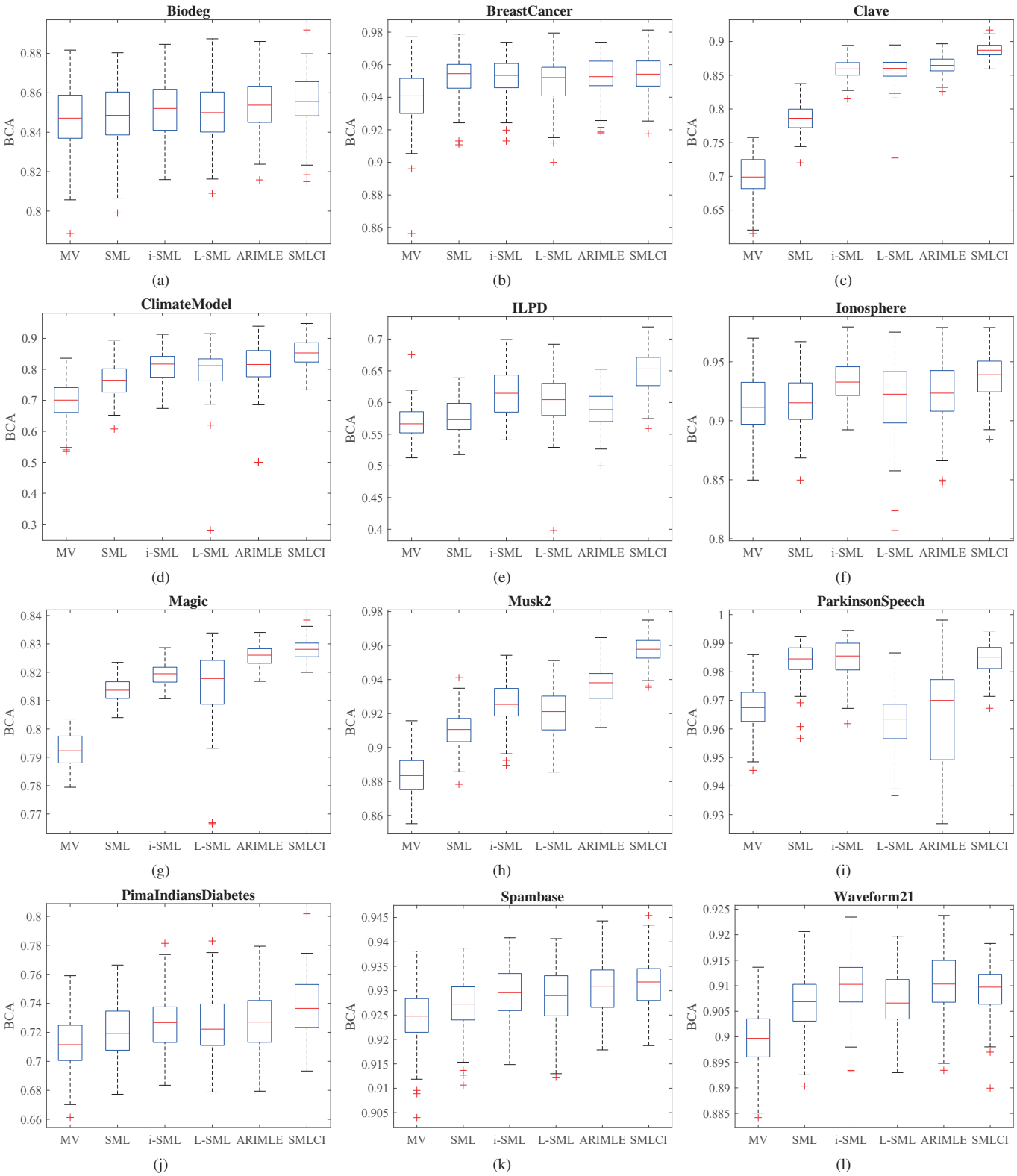

Fig. 2. Performances of the six unsupervised ensemble learning approaches on the 12 UCI datasets.

Unfortunately most real-world applications have more or less class imbalance. In this paper we have proposed an SMLCI approach to explicitly consider the class imbalance. Experiments on 12 UCI datasets verified that SMLCI significantly outperformed the individual base classifiers, and also five existing unsupervised ensemble learning approaches, when the balanced classification accuracy is used as the performance measure.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China (61873321).

REFERENCES

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, pp. 289–300, 1995.

[2] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of theminority class in boosting," in *Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, Sep. 2003, pp. 107–119.

[3] O. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, pp. 62–64, 1961.

[4] ——, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, pp. 214–252, 1964.

[5] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 7th Int'l Joint Conf. of Artificial Intelligence*, Seattle, WA, August 2001, pp. 973–978.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 28, pp. 337–407, 2000.

[7] S. Hashem, "Optimal linear combinations of neural networks," *Neural Networks*, vol. 10, no. 4, pp. 599–614, 1997.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[9] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised ensemble learning with dependent classifiers," *arXiv: 1510.05830*, 2015.

[10] A. Jaffe, B. Nadler, , and Y. Kluger, "Estimating the accuracies of multiple classifiers without labeled data," in *Proc. 18th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS)*, San Diego, CA, May 2015.

[11] LeoBreiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[12] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[13] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for classimbalance learning," *IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[14] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: Review," *International Journal of Computer Science and Network*, vol. 2, no. 1, 2013.

[15] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, "Ranking and combining multiple predictors without labeled data," *Proc. National Academy of Science*, vol. 111, no. 4, pp. 1253–1258, 2014.

[16] E. A. Platanios, A. Blum, and T. M. Mitchell, "Estimating Accuracy from Unlabeled Data," in *Proc. Int'l. Conf. on Uncertainty in Artificial Intelligence (UAI)*, Quebec, Canada, July 2014, pp. 1–10.

[17] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1992.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors," *Nature*, vol. 323, pp. 533–536, 1986.

[19] V. Vapnik, *Statistical Learning Theory*. New York, NY: Wiley Press, 1998.

[20] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Agreement rate initialized maximum likelihood estimator (ARIMLE) for ensemble classifier aggregation and its application in brain-computer interface," in *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, Budapest, Hungary, October 2016, pp. 724–729.

[21] ——, "Spectral meta-learner for regression (SMLR) model aggregation: Towards calibrationless brain-computer interface (BCI)," in *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, Budapest, Hungary, October 2016, pp. 743–749.

[22] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. Boca Raton, FL: CRC press, 2012.