# Multiset Feature Learning for Highly Imbalanced Data Classification

**Fei Wu,[1,2,*] Xiao-Yuan Jing,[1,2,*] Shiguang Shan,[3] Wangmeng Zuo,[4] Jing-Yu Yang[5]**

[1] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China

[2] College of Automation, Nanjing University of Posts and Telecommunications, China

[3] Key Lab of Intelligent Information Process of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, China

[4] School of Computer, Harbin Institute of Technology, China

[5] College of Computer Science and Technology, Nanjing University of Science and Technology, China

[*] Corresponding authors: {wufei_8888@, jingxy_2000@}126.com

**Abstract**

With the expansion of data, increasing imbalanced data has emerged. When the imbalance ratio of data is high, most existing imbalanced learning methods decline in classification performance. To address this problem, a few highly imbalanced learning methods have been presented. However, most of them are still sensitive to the high imbalance ratio. This work aims to provide an effective solution for the highly imbalanced data classification problem. We conduct highly imbalanced learning from the perspective of feature learning. We partition the majority class into multiple blocks with each being balanced to the minority class and combine each block with the minority class to construct a balanced sample set. Multiset feature learning (MFL) is performed on these sets to learn discriminant features. We thus propose an uncorrelated cost-sensitive multiset learning (UCML) approach. UCML provides a multiple sets construction strategy, incorporates the cost-sensitive factor into MFL, and designs a weighted uncorrelated constraint to remove the correlation among multiset features. Experiments on five highly imbalanced datasets indicate that: UCML outperforms state-of-the-art imbalanced learning methods.

## Introduction

Data imbalance means the case that one class severely outnumbers another. Usually, the class with more samples is called majority class and the other one is called minority class. When a classical classifier encounters imbalanced data, it tends to favor the majority class samples. The imbalanced data classification problem has attracted much interest from various communities (Li, Wang, and Bryant 2009; Yu et al. 2013; Pan and Zhu 2013; Huang et al. 2016).

Many methods have been addressed to tackle the imbalanced data classification problem (He and Garcia 2009; Hu et al. 2015), and they can be generally categorized into three kinds: (1) Sampling based methods. They employ undersampling or oversampling technique to transform the class-imbalanced dataset into a balanced one (Chawla et al. 2002; Liu et al. 2006). (2) Cost-sensitive learning based methods (Thai-Nghe et al. 2010; Cristiano and Antonio 2013). This kind of methods considers the costs associated with misclassifying samples. (3) Ensemble learning based methods (Galar et al. 2012; Yang et al. 2014). This kind of methods tries to improve the performance of individual classifiers by inducing several classifiers and combining them to obtain a new and more favorable classifier.

Table 1. Properties of highly imbalanced datasets.

| Dataset | Number of majority class samples | Number of minority class samples | Imbalance ratio (IR) |
|---|---|---|---|
| PC1 | 711 | 61 | 11.7:1 |
| Pageblock | 5245 | 231 | 22.7:1 |
| Glass5 | 205 | 9 | 22.8:1 |
| Yeast7 | 1447 | 37 | 39.2:1 |
| Abalone19 | 4142 | 32 | 128.9:1 |

Generally, datasets whose imbalance ratio is higher than 10:1 can be regarded as the highly imbalanced datasets (Fernándeza et al. 2008). Table 1 shows properties of five highly imbalanced datasets derived from various application fields (Menzies et al. 2007; Alcalá-Fdez et al. 2011). We can see that the majority class samples outnumber the minority class samples severely. Ordinary imbalanced learning methods usually decline in classification performance in highly imbalanced classification scenarios.

Recently, a few methods have been addressed to solve the highly imbalanced data classification problem (López et al. 2013). Granular SVMs-repetitive undersampling (GSVM-RU) (Tang et al. 2009), a modification to support vector machines, can minimize the negative effect of information loss while maximizing the positive effect of data cleaning in the undersampling process. Evolutionary undersampling boost (EUSBoost) (Galar et al. 2013) combines boosting algorithm with evolutionary undersampling that can promote the diversity among classifiers. Besides,

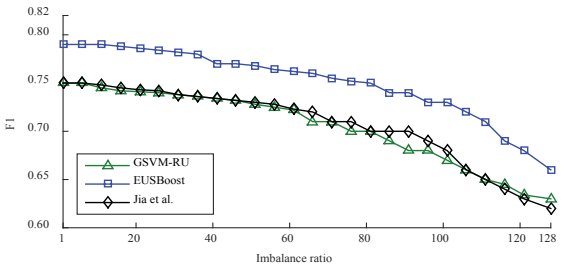

Figure 1. F1 of highly imbalanced learning methods on Abalone19 dataset with increasing imbalance ratio from 1:1 to 128:1.

(Jia et al. 2014) presents two sampling methods based on the borderline synthetic minority over-sampling technique.

## Motivation

Fig. 1 shows the influence of imbalance ratio (IR) to representative highly imbalanced learning methods, i.e., GSVM-RU, EUSBoost and (Jia et al. 2014). Here, we take the Abalone19 dataset as an example and observe the values of the F1 (F-measure with the balance factor $\beta$ being equal to 1) values of these methods with increasing IR (from 1:1 to 128:1). Half of this dataset is taken as the training set and the remainders are used as the testing set. Specifically, we choose half of minority class samples (16 samples) and the same number of samples from the majority class to form the initial training set. In this case, the IR is 1:1. Then, we increase the IR by adding more majority class samples into the initial training set.

We can find that when the IR is increasing to 128:1, F1 of these methods is **much lower than that in initial balanced data scenarios**. Hence, there exists much room for improvement in these methods. Essentially, existing highly imbalanced learning methods **can be classified into one of three kinds of ordinary imbalanced learning methods** mentioned above, and mostly they utilize the sampling and ensemble learning techniques. However, there exist some shortcomings in sampling and ensemble learning techniques based methods, which will be analyzed in Related Work Section. Therefore, highly imbalanced learning is still a challenging task.

## Contribution

The contributions of our study are summarized as following two points:

(1) We intend to **address the highly imbalanced data classification problem from the perspective of feature learning**. Multiset feature learning (MFL) technique can jointly learn features from multiple related sample sets effectively, such that the information of interest is fully exploited. We **are the first to introduce MFL for solving the highly imbalanced data classification problem**.

(2) We provide a multiple sets construction strategy, which can partition the original highly imbalanced data into multiple sets with each holding a class-balanced status. By designing cost-sensitive between-class scatter, we incorporate the cost-sensitive factor into MFL. In addition, we design a weighted uncorrelated constraint to remove the correlation among features learned from different sets.

We call the proposed highly imbalanced learning approach as uncorrelated cost-sensitive multiset learning (UCML). Experiments on five challenging datasets from various fields demonstrate the effectiveness of UCML.

# Related Work

## Class-imbalanced Learning Methods

*A. Sampling technique based methods.* Undersampling based methods balance the distributions between majority class and minority class samples by reducing the majority class samples. Oversampling based methods, however, add the minority class samples to the imbalanced dataset. Majority weighted minority oversampling technique (MWMOTE) (Barua et al. 2014) is a synthetic minority oversampling based method, which generates the synthetic samples by using the weighted informative minority class samples. These methods need to append or remove considerable samples for classifying the highly imbalanced data.

*B. Cost-sensitive learning based methods.* Cost-sensitive multilayer perceptron (CSMLP) (Cristiano and Antonio 2013) is a recently presented algorithm, which uses a single cost parameter to distinguish the importance of class errors. For cost-sensitive learning based methods, how to determine a cost representation is still an important and open problem.

*C. Ensemble learning based methods.* Usually, the ensemble learning based methods are combined with the data sampling technique to address the data imbalance problem (Yang et al. 2014). Undersampling based online bagging with adaptive weight adjustment (WEOB2) (Wang et al. 2015) can adjust the learning bias from majority to minority class effectively with adaptive weight adjustment. (Dubey et al. 2014) presents an ensemble system that combines feature selection algorithm, data sampling technique and binary prediction model. These methods usually focus on the classifier level issue. However, how to effectively guarantee and utilize the diversity of classification ensembles is still an open problem.

The introduction and analysis of highly imbalanced learning methods have been given in Introduction section.

## Multiset Feature Learning (MFL) Methods

The idea of multiset feature learning (MFL) is to jointly learn features from multiple related sample sets, such that the information of interest can be fully exploited (Memisevic et al. 2012). Multiset canonical correlation analysis (MCCA) (Li et al., 2009) exploits the correlation features

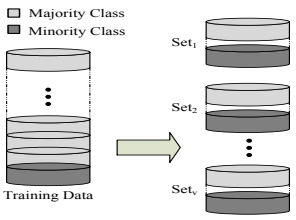


Figure 2. Illustration of multiple sets construction strategy.

among multiple sets. Discriminant analysis based MFL is an important research direction in this domain, including multi-view Fisher discriminant analysis (MFDA) (Diethe et al. 2008) and multi-view discriminant analysis (MvDA) (Kan et al. 2016). MvDA can maximize the between-class variations and minimize the within-class variations of samples in the learning common space from both intra-view and inter-view. To our knowledge, MFL has not been used to solve the imbalanced data classification problem.

## Our Approach

### Multiple Sets Construction Strategy

Fig. 2 illustrates the construction procedure of multiple sets, which includes two steps:

**Step 1:** We randomly partition the majority class samples into multiple blocks, whose number of samples is the same as that of minority class samples.

Since the number of majority class samples might not be exactly in proportion to the number of minority class samples, some majority class samples may be left when multiple blocks have been obtained. We call these left samples as "redundant samples". We delete redundant samples (if the redundant samples are less than half of the minority class samples) or add adequate number of samples copied from original majority class samples (if the redundant samples are more than half of the minority class samples) to construct integral number of blocks. It is noted that the added samples are all randomly copied from original majority class samples.

**Step 2:** We combine each block of majority class with the minority class to form one balanced set. Then, we can obtain multiple balanced sets.

With the designed multiple sets construction strategy, the highly imbalanced data can be transformed to multiple sets. And the highly imbalanced data classification problem can be addressed by using the MFL techniques.

### Cost-sensitive Multiset Feature Learning

Assume that $v$ sets have been constructed. To boost the misclassifying cost and improve the classification performance, we incorporate the cost-sensitive factor into MFL. Concretely, the cost-sensitive factor is embodied in the between-class scatter since it represents the relationship between the majority and minority classes.

Table 2. Cost matrix for UCML.

| | Classified as majority class | Classified as minority class |
|---|---|---|
| Actually majority class | 0 | $cost(1,2)$ |
| Actually minority class | $cost(2,1)$ | 0 |

Let $\mathrm{X}^{(j)} = \left\{ x_{ijk} \middle| i = 1,...,c; k = 1,...,n_{ij} \right\}$ be the samples from the $j^{th}$ set, where $x_{ijk}$ is the $k^{th}$ sample from the $j^{th}$ set of the $i^{th}$ class, $c$ is the number of classes and $n_{ij}$ is the number of samples from the $j^{th}$ set of the $i^{th}$ class. Samples from $v$ sets can be projected to a common space by using the $v$ linear transformations $w_1, w_2, ..., w_v$ denoted as $\Upsilon = \left\{ y_{ijk} = w_j^T x_{ijk} \middle| i = 1,2; j = 1,...,v; k = 1,...,n_{ij} \right\}$. In this common space, the between-class variation $S_B^y$ from all sets is maximized while the within-class variation $S_W^y$ from all sets is minimized. To simplify exposition and ensure clarity, we assume the desired dimension of projected samples equal to one, namely $w_1, w_2, ..., w_v$ are a set of projection vectors. We use $\mu_i = (1/n_i) \sum_{j=1}^{v} \sum_{k=1}^{n_{ij}} y_{ijk}$ as mean of samples in the projected space from the $i^{th}$ class. Here, $n_i$ is the number of samples in the $i^{th}$ class. Then the within-class scatter $S_W^y$ is defined as:

$$S_W^y = \sum_{i=1}^{c} \sum_{j=1}^{v} \sum_{k=1}^{n_{ij}} \left( y_{ijk} - \mu_i \right) \left( y_{ijk} - \mu_i \right)^T . \quad (1)$$

Formally, the within-class scatter in (1) can be reformulated as follows (please refer to the (Kan et al. 2016) for the detailed derivation):

$$S_W^y = \begin{bmatrix} w_1^T w_2^T ... w_v^T \end{bmatrix} \begin{pmatrix} S_{11} & S_{12} & ... & S_{1v} \\ S_{21} & S_{22} & ... & S_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ S_{v1} & S_{v2} & ... & S_{vv} \end{pmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_v \end{bmatrix} = w^T S w , \quad (2)$$

where $w = \left[ w_1^T, w_2^T, ..., w_v^T \right]^T$. $S_{jm}$ in $S$ is defined as follows with $\mu_{ij}^{(x)} = \left( 1/n_{ij} \right) \sum_{k=1}^{n_{ij}} x_{ijk}$ :

$$S_{jm} = \begin{cases} \sum_{i=1}^{c} \left( \sum_{k=1}^{n_{ij}} x_{ijk} x_{ijk}^T - \frac{n_{ij} n_{ij}}{n_i} \mu_{ij}^{(x)} \mu_{ij}^{(x)T} \right) & j = m \\ \sum_{i=1}^{c} \left( - \frac{n_{ij} n_{ij}}{n_i} \mu_{ij}^{(x)} \mu_{ij}^{(x)T} \right) & otherwise \end{cases} . \quad (3)$$

Assume that $cost(1,2)$ denotes the punishment when a majority class sample is misclassified as the minority class, and $cost(2,1)$ means the punishment when a minority-class sample is misclassified as the majority class, as shown in Table 2. We incorporate the cost-sensitive factors $cost(1,2)$ and $cost(2,1)$ in between-class scatter to increase the punishment when minority class samples are misclassified as

the majority class samples. As a result, cost-sensitive MFL makes the classification incline to classify the samples into the minority class. Then the cost-sensitive between-class scatter $S_B^y$ is defined as

$$S_B^y = \sum_{i=1}^{c}\sum_{l=1,l\neq i}^{c} cost(i,l)\left(\mu_i - \mu_l\right)\left(\mu_i - \mu_l\right)^T , \quad (4)$$

where $\mu_l$ is the mean of samples in the projected space from the $l^{th}$ class.

The cost-sensitive between-class scatter can be further reformulated as follows:

$$S_B^y = \left[w_1^T w_2^T ... w_v^T\right]\begin{pmatrix} D_{11} & D_{12} & ... & D_{1v} \\ D_{21} & D_{22} & ... & D_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ D_{v1} & D_{v2} & ... & D_{vv} \end{pmatrix}\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_v \end{bmatrix} = w^T D w . \quad (5)$$

Here, $D_{jm}$ in matrix $D$ is defined as:

$$D_{jm} = \sum_{i=1}^{c}\sum_{\substack{l=1 \\ l\neq i}}^{c} cost(i,l)\left(\frac{n_{ij}n_{im}}{n_i^2}\mu_{ij}^{(x)}\mu_{im}^{(x)T} - \frac{n_{ij}n_{lm}}{n_i n_l}\mu_{ij}^{(x)}\mu_{lm}^{(x)T} - \frac{n_{lj}n_{im}}{n_i n_l}\mu_{lj}^{(x)}\mu_{im}^{(x)T} + \frac{n_{lj}n_{lm}}{n_l^2}\mu_{lj}^{(x)}\mu_{lm}^{(x)T}\right), \quad (6)$$

where $n_{im}$ is the number of samples from the $m^{th}$ set of the $i^{th}$ class, $n_{lm}$ and $n_{lj}$ are separately the numbers of samples from the $m^{th}$ and $j^{th}$ set of the $l^{th}$ class. $\mu_{im}^{(x)} = \left(1/n_{im}\right)\sum_{k=1}^{n_{im}} x_{imk}$ is the mean sample in the $m^{th}$ set of the $i^{th}$ class, and $\mu_{lm}^{(x)}$ and $\mu_{lj}^{(x)}$ are separately the means of samples from the $m^{th}$ and $j^{th}$ set of the $l^{th}$ class.

## Weighted Uncorrelated Constraint

For constructing multiple sets, we partition majority class samples into multiple blocks and combine each block with the minority class to construct a set. Since the minority class is shared by different sets and the samples other than the minority class in different sets are all from the majority class, there may exist correlation among multiple sets. And the correlation in original sets will lead to the correlation in the learned features from multiple sets. Therefore, we consider reducing the adverse correlation in the MFL process.

There already exist efforts to make features learned from single set uncorrelated, including uncorrelated optimal discrimination vectors (UODV) (Jing et al. 2011) and weighted global uncorrelated discriminant transforms (WGUDT) (Jing et al. 2011). UODV and WGUDT separately make features from single set statistically uncorrelated or weighted global uncorrelated, and achieve impressive effects. Inspired by these single-set-based methods, we design a weighted uncorrelated constraint to reduce the statistical correlation among features from multiple sets. The weighted correlation among multiple sets is defined as:

$$cor_{jm} = \sum_{j=1}^{v}\sum_{\substack{m=1 \\ m\neq j}}^{v} \frac{w_j^T S_t^{j,m} w_m}{\sqrt{w_j^T S_t^{j,j} w_j}\sqrt{w_m^T S_t^{m,m} w_m}}, \quad (7)$$

where $S_t^{j,m} = \left(1/N\right)\sum_{q=1}^{N}\sum_{p=1}^{N}\left(x_{jp} - \tilde{\mu}_p^j\right)\left(x_{mq} - \tilde{\mu}_q^m\right)^T$ . $S_t^{j,j}$ and $S_t^{m,m}$ can be computed in the similar way. Here, $x_{jp}$ and $x_{mq}$ denote the $p^{th}$ sample of the $j^{th}$ set and the $q^{th}$ sample of the $m^{th}$ set, respectively. $\tilde{\mu}_p^j$ and $\tilde{\mu}_q^m$ separately denote the **weighted mean sample** corresponding to each sample $x_{jp}$ and $x_{mq}$. $\tilde{\mu}_p^j$ can be calculated by

$$\tilde{\mu}_p^j = \left(1/N\right)\sum_{q=1}^{N}\alpha_{pq}x_{jp},\ \alpha_{pq} = \exp\left(-\left\|x_{jp} - x_{jq}\right\|^2 / 2\sigma^2\right),$$

where $\sigma$ is a scalar constant. Then, our designed uncorrelated constraint is defined as:

$$\sum_{j=1}^{v}\sum_{m=1,m\neq j}^{v} w_j^T S_t^{j,m} w_m = w^T H w = 0 , \quad (8)$$

where $H = \begin{bmatrix} 0 & S_t^{1,2} & \cdots & S_t^{1,v} \\ S_t^{2,1} & 0 & \cdots & S_t^{2,v} \\ \vdots & \vdots & \vdots & \vdots \\ S_t^{v,1} & S_t^{v,2} & \cdots & 0 \end{bmatrix}$. It is noted that since different sets contain the same number of samples, in this part, we use $N$ to denote the number of samples in each set.

## Objective Function and Solution

By combining the multiset within-class scatter, the multiset cost-sensitive between-class scatter and the weighted uncorrelated constraint, we define the objective function of UCML as follows:

$$\max_{w} w^T\left(D - S\right)w \quad s.t.\ w^T H w = 0 . \quad (9)$$

Like in (Jing et al. 2015), the solution of (9) can be obtained by solving the following eigen-equation problem:

$$\left(D - S\right)w = \lambda H w . \quad (10)$$

Once the eigenvectors $w^k\ (k = 1,2,\cdots,d)$ associated with $d$ largest eigenvalues of $\left(D - S\right)^{-1}H$ are obtained, we get $w_1^k, w_2^k, \cdots, w_v^k$ from $w^k$. Let $W_j = \left[w_j^1, w_j^2, ..., w_j^d\right]$ where $j = 1,2,...,v$ and T denote the testing sample set. We can obtain the projected features of training sample set $Z_X^j$ and testing sample set $Z_T^j$ separately by $Z_X^j = W_j^T \mathrm{X}^{(j)}$ and $Z_T^j = W_j^T \mathrm{T}$ for $j = 1,2,...,v$.

For the $j^{th}\left(j = 1,2,...,v\right)$ set, we firstly use the nearest neighbor (NN) classifier with the cosine distance to classify $Z_T^j$ on $Z_X^j$. Then we can obtain $v$ predicted results for each testing sample in T. Next, we can adopt the majority voting strategy to make final decision for each test sample.

## Complexity Analysis

The time cost of UCML mainly includes two parts: (1) calculating matrices $S$, $D$, and $H$; (2) solving the generalized eigenvalue problem in (10). Specifically, matrices

calculation needs the time complexity of $v^2\left(N+c^2+N^2\right)(\dim)^2$, where dim denotes the dimensionality of samples. Solving the generalized eigenvalue problem in (10) needs the time cost of $(v\dim)^3$. Therefore, our approach needs the time cost of $v^2\left(N+c^2+N^2\right)(\dim)^2+(v\dim)^3$.

# Experiments

## Competing Methods

In the experiment, we compare our UCML approach with state-of-the-art related methods including: highly imbalanced learning methods: **GSVM-RU** (Tang et al. 2009), **EUSBoost** (Galar et al. 2013) and (**Jia et al.** 2014); representative ordinary imbalanced learning methods: **MWMOTE** (Barua et al. 2014), **CSMLP** (Cristiano and Antonio 2013), **WEOB2** (Wang et al. 2015) and (**Dubey et al.** 2014); and representative multiset learning methods: **MCCA** (Li et al. 2009) and **MvDA** (Kan et al. 2016).

## Evaluation Measures and Experimental Setting

We employ three commonly used measures, including Precision, Recall, and F-measure, to evaluate the performances. Assume that $A$, $B$, $C$ and $D$ are the number of minority class samples that are classified as minority class, the number of minority class samples that are classified as majority class, the number of majority class samples that are classified as minority class, and the number of majority class samples that are classified as majority class, respectively, these measures can be calculated as:

(1) $Precision = A/(A+C)$.

(2) $Recall = A/(A+B)$.

(3) $F\text{-}measure = \left(1+\beta^2\right)\dfrac{Precision \times Recall}{\left(\beta^2 \times Precision\right)+Recall}$. It is a trade-off between the Precision and Recall. A greater value for $\beta$ indicates the higher importance of recall over precision. In this paper, we use the widely used F1, that is F-measure with $\beta=1$. In addition, we also evaluate the class-imbalanced learning performance of our approach by using F2 (F-measure with $\beta=2$), like in (Maratea et al. 2014).

Obviously, an ideal method should hold high values of Precision, Recall, F1 and F2. In experiments, we randomly select 50% samples to construct the training set for all datasets, and use the remained samples for testing. We repeat random selection 20 times and record the average results. Assume that the first class is the majority class and the second class is the minority class. Then $cost(1,2)$ and $cost(2,1)$ are separately set as $cost(1,2)=1$ and $cost(2,1)=$ rounded value of $N_1/N_2$, where $N_1$ and $N_2$ denote the numbers of majority and minority class samples.

Table 3. Experimental results on the PC1 dataset.

| Method | Precision | Recall | F1 | F2 |
|---|---|---|---|---|
| MCCA | 0.42±0.03 | 0.48±0.05 | 0.44±0.04 | 0.47±0.05 |
| MvDA | 0.46±0.06 | 0.51±0.04 | 0.47±0.04 | 0.50±0.06 |
| MWMOTE | 0.72±0.04 | 0.83±0.05 | 0.77±0.04 | 0.81±0.05 |
| CSMLP | 0.66±0.03 | 0.84±0.02 | 0.73±0.02 | 0.80±0.04 |
| WEOB2 | 0.69±0.02 | 0.82±0.02 | 0.76±0.02 | 0.79±0.04 |
| Dubey et al. | 0.71±0.04 | 0.80±0.03 | 0.75±0.02 | 0.78±0.03 |
| GSVM-RU | 0.73±0.03 | 0.82±0.04 | 0.76±0.03 | 0.80±0.06 |
| EUSBoost | 0.72±0.04 | 0.81±0.02 | 0.75±0.03 | 0.79±0.05 |
| Jia et al. | 0.72±0.03 | 0.80±0.03 | 0.76±0.03 | 0.78±0.05 |
| UCML | **0.74**±0.03 | **0.84**±0.04 | **0.78**±0.02 | **0.82**±0.03 |

Table 4. Experimental results on the Pageblock dataset.

| Method | Precision | Recall | F1 | F2 |
|---|---|---|---|---|
| MCCA | 0.39±0.06 | 0.45±0.04 | 0.42±0.05 | 0.44±0.02 |
| MvDA | 0.44±0.04 | 0.47±0.05 | 0.45±0.05 | 0.46±0.03 |
| MWMOTE | 0.88±0.05 | 0.90±0.05 | 0.88±0.04 | 0.90±0.05 |
| CSMLP | 0.89±0.06 | 0.92±0.02 | 0.89±0.03 | 0.91±0.02 |
| WEOB2 | 0.91±0.03 | 0.93±0.03 | 0.91±0.03 | 0.93±0.03 |
| Dubey et al. | 0.88±0.02 | 0.90±0.02 | 0.89±0.03 | 0.90±0.01 |
| GSVM-RU | 0.89±0.03 | 0.92±0.04 | 0.90±0.04 | 0.91±0.03 |
| EUSBoost | 0.92±0.02 | 0.93±0.03 | 0.92±0.02 | 0.93±0.04 |
| Jia et al. | 0.89±0.05 | 0.88±0.04 | 0.88±0.05 | 0.88±0.02 |
| UCML | **0.93**±0.03 | **0.94**±0.03 | **0.94**±0.04 | **0.94**±0.02 |

Table 5. Experimental results on the Glass5 dataset.

| Method | Precision | Recall | F1 | F2 |
|---|---|---|---|---|
| MCCA | 0.43±0.12 | 0.76±0.11 | 0.56±0.10 | 0.66±0.08 |
| MvDA | 0.45±0.09 | 0.75±0.08 | 0.54±0.08 | 0.66±0.06 |
| MWMOTE | 0.71±0.05 | 0.82±0.03 | 0.74±0.04 | 0.80±0.04 |
| CSMLP | 0.72±0.03 | 0.83±0.04 | 0.75±0.02 | 0.81±0.05 |
| WEOB2 | 0.76±0.06 | 0.83±0.07 | 0.78±0.04 | 0.81±0.07 |
| Dubey et al. | 0.77±0.04 | 0.86±0.03 | 0.82±0.03 | 0.84±0.03 |
| GSVM-RU | 0.78±0.04 | 0.88±0.03 | 0.82±0.03 | 0.86±0.02 |
| EUSBoost | 0.78±0.05 | 0.85±0.04 | 0.82±0.04 | 0.84±0.05 |
| Jia et al. | 0.74±0.05 | 0.79±0.05 | 0.77±0.05 | 0.78±0.04 |
| UCML | **0.81**±0.02 | **0.89**±0.04 | **0.86**±0.03 | **0.87**±0.06 |

The parameter $\sigma^2$ in the weighted uncorrelated constraint is set by using 5-fold cross validation on the training set.

## Software Defect Prediction Application

To validate the effectiveness of UCML for software defect prediction, we conduct experiments on the PC1 dataset (Menzies et al. 2007). Each sample in this dataset has 38 features. Table 3 shows the experimental results on PC1. We can see that UCML can achieve better results.

## Document Classification Application

To validate the effectiveness of UCML for document classification, we conduct experiments on the Pageblock[1] dataset. The imbalance ratio is 22.7:1 and each sample has 10 features. Table 4 shows the experimental results. We can see that UCML obtains the best classification results.

## Object Classification Application

Object classification also usually encounters the highly imbalanced data. Thus, we conduct experiment on this type of dataset like Glass5 (Alcalá-Fdez et al. 2011). Each sample in this dataset has 9 features and the imbalance ratio is 22.8:1. Table 5 shows the experimental results. We can see that our UCML is superior to other compared methods.

[1] UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/, 2009.

## Bio-information Prediction Application

We conduct experiments on the Yeast7 and Abalone19 datasets (Alcalá-Fdez et al. 2011) for bio-information prediction. Yeast7 is usually used to predict the cellular localization sites of proteins. Abalone19 is usually used to predict the age of abalone. Each sample has 8 features in these two datasets.

Tables 6 and 7 separately show the prediction results on Yeast7 and Abalone19. UCML obtains the best prediction results on both datasets. We also conduct the statistical test (Draper et al. 2002) to analyze the results in Tables 3-7. The test results indicate that UCML makes a statistically significant difference in comparison with other methods.

Table 6. Experimental results on the Yeast7 dataset.

| Method | Precision | Recall | F1 | F2 |
|---|---|---|---|---|
| MCCA | 0.51±0.05 | 0.37±0.02 | 0.44±0.05 | 0.39±0.03 |
| MvDA | 0.54±0.03 | 0.39±0.02 | 0.45±0.03 | 0.41±0.05 |
| MWMOTE | 0.67±0.04 | 0.56±0.03 | 0.60±0.04 | 0.58±0.07 |
| CSMLP | 0.71±0.03 | 0.57±0.02 | 0.63±0.04 | 0.59±0.08 |
| WEOB2 | 0.70±0.02 | 0.59±0.03 | 0.65±0.05 | 0.61±0.05 |
| Dubey et al. | 0.73±0.04 | 0.60±0.02 | 0.63±0.04 | 0.62±0.07 |
| GSVM-RU | 0.73±0.03 | 0.61±0.04 | 0.67±0.03 | 0.63±0.04 |
| EUSBoost | 0.77±0.02 | 0.61±0.02 | 0.68±0.05 | 0.64±0.06 |
| Jia et al. | 0.72±0.04 | 0.60±0.03 | 0.66±0.05 | 0.62±0.07 |
| UCML | **0.77**±0.03 | **0.65**±0.04 | **0.71**±0.03 | **0.67**±0.02 |

Table 7. Experimental results on the Abalone19 dataset.

| Method | Precision | Recall | F1 | F2 |
|---|---|---|---|---|
| MCCA | 0.33±0.04 | 0.52±0.04 | 0.41±0.03 | 0.47±0.05 |
| MvDA | 0.38±0.03 | 0.53±0.03 | 0.44±0.02 | 0.49±0.03 |
| MWMOTE | 0.53±0.02 | 0.58±0.04 | 0.54±0.03 | 0.57±0.04 |
| CSMLP | 0.51±0.01 | 0.55±0.02 | 0.53±0.02 | 0.54±0.05 |
| WEOB2 | 0.55±0.02 | 0.60±0.03 | 0.57±0.03 | 0.59±0.07 |
| Dubey et al. | 0.58±0.05 | 0.67±0.03 | 0.62±0.04 | 0.65±0.06 |
| GSVM-RU | 0.60±0.03 | 0.70±0.04 | 0.63±0.03 | 0.68±0.04 |
| EUSBoost | 0.63±0.02 | 0.69±0.02 | 0.66±0.03 | 0.68±0.05 |
| Jia et al. | 0.59±0.04 | 0.68±0.02 | 0.62±0.03 | 0.66±0.02 |
| UCML | **0.67**±0.03 | **0.75**±0.05 | **0.71**±0.04 | **0.73**±0.07 |

Table 8. F1 values of $\text{UCML}_{\text{noboth}}$, $\text{UCML}_{\text{nocost}}$, and $\text{UCML}_{\text{nowei}}$.

| Dataset | $\text{UCML}_{\text{noboth}}$ | $\text{UCML}_{\text{nocost}}$ | $\text{UCML}_{\text{nowei}}$ | UCML |
|---|---|---|---|---|
| PC1 | 0.74 | 0.76 | 0.75 | **0.78** |
| Pageblock | 0.91 | 0.93 | 0.92 | **0.94** |
| Glass5 | 0.82 | 0.85 | 0.83 | **0.86** |
| Yeast7 | 0.65 | 0.68 | 0.66 | **0.71** |
| Abalone19 | 0.67 | 0.70 | 0.69 | **0.71** |

## Effectiveness of Important Components

Multiset feature learning (main body of our approach), cost-sensitive factor, and weighted uncorrelated constraint are three important components of our approach. In this subsection, we specially evaluate their effectiveness. We perform our approach without the cost-sensitive factor and weighted uncorrelated constraint, and we call this version as "$\text{UCML}_{\text{noboth}}$". In addition, we perform our approach without the cost-sensitive factor or weighted uncorrelated constraint, which are separately called "$\text{UCML}_{\text{nocost}}$" and "$\text{UCML}_{\text{nowei}}$". The experimental results of $\text{UCML}_{\text{noboth}}$, $\text{UCML}_{\text{nocost}}$, $\text{UCML}_{\text{nowei}}$ and UCML are given in Table 8.

From the table, we can see that the F1 values of $\text{UCML}_{\text{noboth}}$ are obviously lower than those of UCML, but are still comparable to other methods. In addition, $\text{UCML}_{\text{nocost}}$ and $\text{UCML}_{\text{nowei}}$ can improve the results of $\text{UCML}_{\text{noboth}}$. These results demonstrate the effectiveness of the three components in our approach.

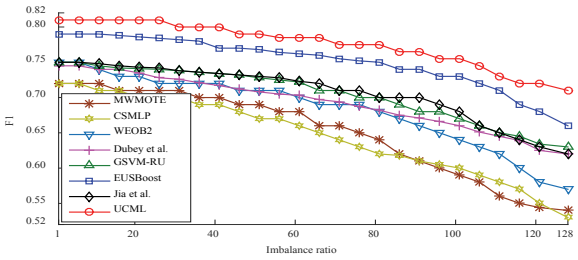


Figure 3. F1 vs. imbalance ratio on Abalone19.

## Evaluation of the Influence of IR to UCML

Fig. 3 illustrates the influence of IR to UCML on Abalone19. The experimental setting can be found in Fig. 1. We can find that when the IR is increasing, F1 values of compared methods decline. When the IR reaches the maximized level (128:1), the F1 values are significantly lower than those corresponding to the IR of 1:1 for all competing methods. For UCML, its F1 experiences a relatively smaller decline, which means that UCML is relatively robust to highly imbalance ratio as compared with related methods.

## Parameter Analysis

For the parameter $\sigma^2$ in our approach, we search the parameter space $\left[2^{-3},2^{-2},2^{-1},2^{0},2^{1},2^{2},2^{3}\right]\times\sigma_0^2$, where $\sigma_0^2$ is the mean square distance of training data. Here, we evaluate the influence of $\sigma^2$ on the prediction result. Fig. 4 shows the F1 values of UCML versus different values of $\sigma^2$ on PC1. We can see that the performances of our approach are stable with respect to $\sigma^2$ in the range of $\left[2\sigma_0^2,4\sigma_0^2\right]$. For simplicity, we set $\sigma^2$ as $2\sigma_0^2$ on PC1. A similar phenomenon also exists on the other datasets.

With respect to the computational time, generally, our approach needs comparable time as compared with MCCA, MvDA, MWMOTE and (Jia et al. 2014), and needs less time than CSMLP, WEOB2, (Dubey et al. 2014), GSVM-RU and EUSBoost.

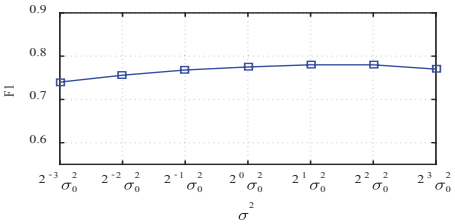


Figure 4: F1 versus $\sigma^2$ on PC1.

## Conclusion

In this paper, we are devoted to addressing the highly imbalanced learning problem from the perspective of feature learning. We propose a novel approach named UCML.

This is the first attempt towards introducing the idea of MFL into imbalanced learning. We conduct experiments on five highly imbalanced datasets from various application fields. The results demonstrate that UCML outperforms state-of-the-art highly imbalanced learning methods. The experimental results indicate that three important components of our approach are effective. We also find that our approach is more robust to high imbalance ratio.

## Acknowledgments

The authors want to thank the editors and anonymous reviewers for their constructive comments and suggestions. The work described in this paper was supported by the National Nature Science Foundation of China under Project Nos. 61272273, 61671182, 61672281 and 61233011, and the Research Project of NJUPT (XJKY14016).

## References

Alcalá-Fdez, J.; Fernandez, A.; Luengo, J.; Derrac, J.; García, S.; Sánchez, L.; and Herrera, F. 2011. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing* 17(2-3): 255-287.

Barua, S.; IsIam, M. M.; Yao, X.; and Murase, K. 2014. MWMOTE—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowledge Data Engineering* 26(2): 405-425.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321-357.

Cristiano, L.; and Antonio, P. B. 2013. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Trans. Neural Networks and Learning Systems* 24(6): 888-899.

Diethe, T.; Hardoon, D. R.; and Shawe-Taylor, J. 2008. Multiview fisher discriminant analysis. *NIPS Workshop on Learning from Multiple Sources*.

Draper, B. A.; Yambor, W. S.; and Beveridge, J. R. 2002. Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures. *Empirical Evaluation Methods in Computer Vision*, 1-15.

Dubey, R.; Zhou, J.; Wang, Y.; Thompson, P. M.; and Ye, J. 2014. Analysis of sampling techniques for imbalanced data: an n=648 ADNI study. *Neuroimage* 87(3): 220-241.

Fernándeza, A.; Garcíaa, S.; Jesusb, M. J. D.; Herrera, F. 2008. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems* 159(18): 2378-2398.

Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; and Herrera, F. 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. System, Man, Cybernetics, Part C* 42(4): 463-484.

Galar, M.; Fernandez, A.; Barrenechea, E.; and Herrera, F. 2013. EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition* 46(12): 3460-3471.

He, H.; and Garcia; E. A. 2009. Learning from imbalanced data. *IEEE Trans. Knowledge Data Engineering* 21(9): 1263-1284.

Hu, J.; Yang, H.; King, I.; Lyu, M. R.; and So, A. M. C. 2015. Kernelized online imbalanced learning with fixed budgets. In *AAAI*, 2666-2672.

Huang, C.; Li, Y.; Change Loy, C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *CVPR*, 5375-5384.

Jia, P. F.; Zhang, C. K.; and Zhang, Z. Y. 2014. A new sampling approach for classification of imbalanced data sets with high density. In *BIGCOMP*, 217-222.

Jing, X.; Li, S.; Zhang, D.; and Yang, J. 2011. Face recognition based on local uncorrelated and weighted global uncorrelated discriminant transforms. In *ICIP*, 3049-3052.

Jing, X. Y.; Liu, Q.; Wu, F.; Xu, B.; Zhu, Y.; and Chen, S. 2015. Web page classification based on uncorrelated semi-supervised intra-view and inter-view manifold discriminant feature extraction. In *IJCAI*, 2255-2261.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-view discriminant analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence* 38(1): 188-194.

Li, Q.; Wang, Y.; and Bryant, S. H. 2009. A novel method for mining highly imbalanced high-throughput screening data in pubchem. *Bioinformatics* 25(24): 3310-3316.

Li, Y. O.; Adali, T.; Wang, W.; and Calhoun, V. D. 2009. Joint blind source separation by multiset canonical correlation analysis. *IEEE Trans. Signal Processing* 57(10): 3918-3929.

Liu, X. Y.; Wu, J.; and Zhou, Z. H. 2006. Exploratory undersampling for class imbalance learning. In *ICDM*, 965-969.

López, V.; Fernández, A.; Del Jesus, M. J.; and Herrera, F. 2013. A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. *Knowledge-Based Systems* 38: 85-104.

Maratea, A.; Petrosino, A.; and Manzo, M. 2014. Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences* 257: 331-341.

Memisevic, R. 2012. On multi-view feature learning. In *ICML*.

Menzies, T.; Greenwald, J.; and Frank, A. 2007. Data mining static code attributes to learn defect predictors. *IEEE Trans. Software Engineering* 33(1): 2-13.

Pan, S.; and Zhu, X. 2013. Graph classification with imbalanced class distributions and noise. In *IJCNN*, 1586-1592.

Tang, Y. C.; Zhang, Y. Q.; Chawla, N. V.; and Krasser, S. 2009. SVMs modeling for highly imbalanced classification. *IEEE Trans. Systems, Man, and Cybernetics Part B* 39(1): 281-288.

Thai-Nghe, N.; Gantner, Z.; and Schmidt-Thieme, L. 2010. Cost-sensitive learning methods for imbalanced data. In *IJCNN*, 1-8.

Wang, S.; Minku, L. L.; and Yao, X. 2015. Resampling-based ensemble methods for online class imbalance learning. *IEEE Trans. Knowledge Data Engineering* 27(5): 1356-1368.

Yang, P. Y.; Yoo, P. D.; Fernando, J.; Zhou, B. B.; Zhang, Z. L.; and Zomaya, A. Y. 2014. Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Trans. Cybernetics* 44(3): 463-484.

Yu, D. J.; Hu, J.; Tang, Z. M.; Shen, H. B.; Yang, J.; and Yang, J. Y. 2013. Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing*, 104: 180-190.