# Cost sensitive Ranking Support Vector Machine for Multi-label Data Learning

Peng Cao[1], Xiaoli Liu[1], Dazhe Zhao[1], Osmar Zaiane[2],

[1]Computer Science and Engineering, Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, China
[2]University of Alberta, Canada
caopeng@cse.neu.edu.cn, liux3795@umn.edu, zhaodz@neusoft.com, zaiane@ualberta.ca

**Abstract.** Multi-label data classification has become an important and active research topic, where the classification algorithm is required to deal with prediction of sets of label indicators for instances simultaneously. Label powerset (LP) method reduces the multi-label classification problem to a single-label multi-class classification problem by treating each distinct combination of labels. However, the predictive performance of LP is challenged with imbalanced distribution among the labelsets, deteriorating the performance of traditional classifiers. In this paper, we study the problem of multi-label imbalanced data classification and propose a novel solution, called CSRankSVM (Cost sensitive Ranking Support Vector Machine), which assigns a different misclassification cost for each labelset to effectively tackle the problem of imbalance for Multi-label data. Empirical studies on popular benchmark datasets with various imbalance ratios of labelsets demonstrate that the proposed CSRankSVM approach can effectively boost classification performances in multi-label datasets.

**Keywords:** Multi-label learning, Imbalanced data, Classification, Rank SVM

## 1 Introduction

In traditional label learning, each object is represented by a single instance and associated with a single label. Typically, binary classifiers are considered where only two classes exist, but in many applications more classes are used and we call these problems multi-class classification problems. Again each instance is associated with a single label. In multi-label classification problems, instances may be associated to more than just one label. Multi-label data classification has a wide variety of real world applications [1,2], e.g. text categorization, scene classification, semantic video, annotation and biological data analysis.

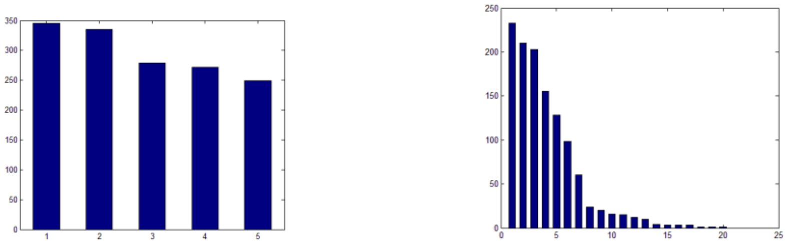
Conventional multi-label learning algorithms aim to find a mapping from the feature space $X \subseteq \mathbb{R}^{d}$ to the label space $Y \subseteq \{0, 1\}^{q}$, wherein $q$ is the number of labels. A simple yet effective multi-label learning method, called label powerset (LP) [3-4], considers each distinct combination of labels that exist in the training set as a different class value of a single-label classification task, where each class denotes a unique $2^{q}$-

dimensional label vector. LP has the advantage of taking label correlations into consideration. However, the resulting distribution of the multiple labelsets (classes) is skewed, since many of these labelsets are usually associated with very few training examples due to a large number of labelsets appearing in the training set. Most multi-label data have hundreds of labels, with each instance being associated with a subset of them. Intuitively, it is easy to see that the more different labels exist, the more possibilities there are, and that some of them have a very low/high presence. For example, assuming an image database (samples in Fig.1) and the task of scene classification, the labelset with the combination of *sunset* and *beach* could be more than the one of the *foliage* and *urban*. Also in textual repositories and document categorisation, documents have many labels and these are unevenly distributed. For the article classification, the labelset with the combination of *politics* and *economic* (e.g. Political risks may foil economic reform in China) could be more than the one of *sport* and *technology* (e.g. Professional sports teams are adopting advanced imagery technology to improve the performance of athletes and their recovery from injuries). The process of learning from imbalanced labelsets usually tends to be overwhelmed by the majority labelset and ignores the minority labelset examples, since most classifiers assume an even distribution of examples among classes and an equal misclassification cost. The imbalanced data issue has been deeply studied for single label classification [5]. This problem also affects multi-label datasets, and the imbalance level in multi-label datasets is much more significant than in binary or multi-class datasets in general.

**(a)** foliage and urban **(b)** sunset and beach

**Fig. 1** The example of imbalanced labelsets in multi-label learning

The issue of imbalanced data in multi-label learning can be tackled from two perspectives: individual label and labelset. In our work, we consider the class imbalance in the labelset space rather than the label space [6,7] since the labelsets consider the high-order correlation among labels, exploiting the relation of multiple labels more effectively and intrinsically. Moreover, even when a dataset is balanced from the view of individual labels, which doesn't require imbalanced data learning, the imbalanced data issue still exists from the view of labelsets, such as the plant dataset(Fig.2.).



(a) The histogram of individual label (b) The histogram of labelsets

**Fig. 2.** The plant dataset

The common methods to solve data imbalance are data re-sampling perspective (i.e. undersampling large classes and oversampling small classes) and algorithm perspective (i.e. treating the imbalance problem in the algorithm itself). The main disadvantage of re-sampling techniques is that they may cause loss of important information or introduce noisy data, since they change the original data distribution, especially for multi-label data with complicated label correlation, so as to resulting in model overfitting. To deal with the issue of imbalanced data in multi-label learning, we propose a novel labelset based multi-label classification method, called Cost sensitive Rank Support Vector Machines (CSRankSVM). The objective of CSRankSVM is to minimize the weighted ranking loss with the weighting scheme with respect to each sample while having a low complexity, so as to be able to learn more characteristics of samples with the minority labelset by setting a high cost to the misclassification of a minority labelset sample without modifying the data distribution.

Our proposed method combines both idea of problem transformation and the algorithm adaptation, both of which consider the label correlation when processing and learning. Firstly, the Label Powerset transforms the original data into multiple labelsets. This is problem transformation. Then, the CSRankSVM models the classifier by considering the imbalanced labelsets generated through assigning a different misclassification cost for relevant labelset of each instance. This is the algorithm adaptation. The contributions of this work can be listed as follows:

1) First, we define a new metric to assess the imbalance level of labelsets generated by LP in the multi-label data, and propose two solutions to solve the imbalanced issue of labelsets, CSRankSVM and CSRankSVM-***p***.
2) We empirically demonstrate that it improves the traditional RankSVM[9], and outperforms the state-of-the-art approaches for dealing with multi-label data on six benchmark datasets in terms of macro-FM, micro-FM and ranking loss.

## 2 Imbalanced labelsets in multi-label data learning

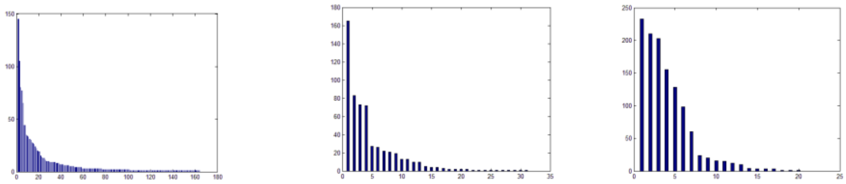
We formally define the multi-label classification problem as this: Let $X$ denotes the space of instances and $Y=\left\{y_1,...,y_q\right\}$ denotes the class labels where $|Y|=q$ . $T=\left\{(x_1,Y_1),...,(x_p,Y_p)\right\}$ $(|T|=p)$ is the multi-label training dataset. $Y_i \subset Y \subset 2^q$ is a labelset identity associated with instance $x_i \in X$ , and the set $Y'$ ($Y_j$, $j$=1,…,|$Y$|) is used to denote the whole finite set labelset appeared in the training set. The goal of the multi-label classification is to get a classifier $h: X \to 2^q$ that generalizes well on both these training instances and unseen ones in the sense of optimizing some expected risk function with respect to a specific empirical loss function.

For binary imbalanced data, the class with fewer instances is the minority class, and the other class is the majority class. The imbalance level is easily measured taking into account only the two classes. However, for labelsets in multi-label data, there exist multiple majority labelsets (classes) and multiple minority labelsets (classes), therefore the imbalance is more challenging. Several interesting research questions are raised here: How to measure the imbalance level of labelsets for multi-label data? How to overcome the issue of imbalanced labelset learning in multi-label data? Is the imbalanced data learning method still effective for multi-label data? Can the classification performance of multi-label data be increased through tackling the imbalanced labelsets by the scheme of cost sensitive learning?

In traditional binary imbalanced data, the imbalanced data level between labels is assessed by the ratio between minority and majority examples. In this scenario of data with multi-labels, we need to introduce a new measure to assess the imbalance level of the whole dataset considering all the labels. We define the level of imbalance of labelsets, named *ImbalR*, which indicates a peakedness level of the histogram distribution of labelsets value in descending order, inspired by the idea of kurtosis.

$$ImbalR = \frac{\sum_{i=1}^{l}(Y_i - Y_{max})^4}{(l-1)s^4} \tag{1}$$

Where $s = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(Y_i - Y_{max})}$, $Y_i$ is the amount of instances in the *i-th* labelset, $Y_{max}$ is the amount of instances of the labelset with the maximum amount of instances, $l$ is the number of all the labelsets. If the histogram is more peaked (or inversely flatter), the distribution of labelsets value is more (or less) imbalanced, then *ImbalR* is larger (or smaller). The histogram of labelsets and *ImbalR* of multi-label datasets is shown in Fig.3.



(a) Yeast: *ImbalR=1903* (b) Plant: *ImbalR=381* (c) Image: *ImbalR=43.4*

**Fig. 3.** The *ImbalR* of multi-label datasets

## 3 Cost sensitive RankSVM

Current cost-sensitive learning research has been focused on binary or multi-class classification, but never yet on multi-label classification. Rank Support Vector Machine (RankSVM) [8] is an excellent kernel-based tool for multi label classification. However, the conventional SVM based method performs poorly on imbalanced learning because it pays less attention to the minority class [9-10]. In order to overcome the imbalanced labelsets of multi-label classification, we inject the idea of cost sensitive learning into the RankSVM algorithm by coupling a cost $\lambda_i$ to each instance. Therefore, we define the quadratic optimization problem of cost sensitive RankSVM as follows:

$$\min \frac{1}{2}\sum_{k=1}^{q}\|w_k\|^2 + C\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}\sum_{(m,n)\in Y_i\times\bar{Y}_i}\lambda_i\xi_{imn}, \tag{2}$$
$$\text{s.t. } < w_m - w_n, x_i > + b_m - b_n \geq 1 - \xi_{imn}, \quad \xi_{imn} \geq 0, (m,n)\in(Y_i\times\bar{Y}), i = 1,\ldots p.$$

Dealing with imbalanced labelsets, cost-sensitive learning scheme assumes higher misclassification costs with minority labelset. The misclassification costs play a crucial role in the construction of a cost sensitive learning model for achieving expected classification results [10]. The most straightforward solution is to automatically generate the misclassification cost vector in accordance with the labelset distribution, which usually is in the form of a weight scheme inversely proportional to the number of samples in the labelsets. The relevant and minority labelsets of each instance are associated with higher misclassification cost values. Therefore, it is necessary to propose a RankSVM with cost sensitive learning to resolve the multi-label datasets.

The Lagrangian for the primal form in (2) can be expressed as (dual variables $\alpha_{imn}$ and $\eta_{imn}$ related to the constraints of (2)):

$$L = \frac{1}{2}\sum_{k=1}^{q}\|w_k\|^2 + C\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}\sum_{(m,n)\in Y_i\times\bar{Y}_i}\lambda_i\xi_{imn}$$
$$-\sum_{i=1}^{p}\sum_{(m,n)\in Y_i\times\bar{Y}_i}\alpha_{imn}(< w_m - w_n, x_i > + b_m - b_n - 1 + \xi_{imn}) - \sum_{i=1}^{p}\sum_{(m,n)\in Y_i\times\bar{Y}_i}\eta_{imn}\xi_{imn} \tag{3}$$

After some resolving according to the Karush-Kuhn-Tucker (KKT) conditions, we can get the following equations:

$$\partial_{w_k} L = w_k - \sum_{i=1}^{p}\left(\sum_{(m,n)\in Y_i\times\bar{Y}_i} c_{imn}\alpha_{imn}\right)x_i = 0; \tag{4}$$

$$\partial_{b_k} L = \sum_{i=1}^{p}\sum_{(m,n)\in Y_i\times\bar{Y}_i} c_{imn}\alpha_{imn} = 0; \;\; \partial_{\xi_{imn}} L = \frac{C\lambda_i}{|Y_i||\bar{Y}_i|} - \alpha_{imn} - \eta_{imn} = 0 \tag{5}$$

Where

$$c_{imn} = \begin{cases} 0 & \text{if } m \neq k \text{ and } n \neq k \\ +1 & \text{if } m = k \\ -1 & \text{if } n = k \end{cases} \tag{6}$$

By introducing (4)-(6) into the Lagrangian (3), the dual of the optimization problem can be expressed as:

$$\max_{\alpha_{imn}} W(\alpha) = -\frac{1}{2}\sum_{k=1}^{q}\sum_{i,j=1}^{p}\beta_{ki}\beta_{kj} < x_i, x_j > + \sum_{i=1}^{p}\sum_{(m,n)\in Y_i\times\bar{Y}_i}\alpha_{imn},$$
$$\text{s.t. } \sum_{i=1}^{p}\sum_{(m,n)\in Y_i\times\bar{Y}_i} c_{imn}\alpha_{imn} = 0, \alpha_{imn} \in [0, \frac{C\lambda_i}{|Y_i||\bar{Y}_i|}], (m,n)\in(Y_i\times\bar{Y}), i = 1,\ldots p. \tag{7}$$

where $\beta_{ki} = \sum_{(m,n)\in Y_i\times\bar{Y}_i} c_{imn}\alpha_{imn}, k = 1,\ldots q$

Then the Franke and Wolfe's method [8] is applied to solve the optimization problem, as done in the traditional RankSVM. Additionally, in order to reduce the excessive labelsets produced by LP, a Pruned Problem Transformation method (PPT) [11] is applied to prune the labelsets with a threshold (line 3—6), and the final pruned labelsets, labelsets-***p*** are obtained. The detailed algorithm CRankSVM (CRankSVM-***p***) is shown in Algorithm 1.

**Algorithm 1:** CSRankSVM (CSRankSVM-***p***)

**Input:** Training set $D$; Test set $T$; threshold parameter of pruning $t$;

1. Transform the original dataset $D$ by LP into labelsets $L$
2. Compute the amount of instance in each labelset
3. **if** pruning the labelsets is required /* case for CSRankSVM-***p*** */
    **for** $j$=1 to $|L|$
4. Split the $L_j$ into multi disjoint subsets where subsets occur more than $t$ times in the $D$
5. The instances in $L_j$ is duplicated and assigned one of the subsets
6. The original labelset $L_j$ is discarded
    **end for**
    **end if**
    **for** $i$=1 to $|D|$
7. Obtain the relevant LabelSet $L(x_i)$ and irrelevant LabelSet $\overline{L(x_i)} = L \setminus L(x_i)$
8. Calculate the weight $\lambda_i$ of $x_i$, $\lambda_i$ =tansig($|L(x_i)| / |\overline{L(x_i)}|$)
    **end for**
9. Optimize CSRankSVM with weight vector $\boldsymbol{\lambda}$ according to Equations 3-7

# 4 Experimental study

## 4.1 Experimental setting

To evaluate the performance of our proposed method in multi-label classification tasks, a total of six common multi-label datasets are used in this study. **Table 1** shows some useful statistics of these datasets, such as the number of instances in the training and test sets, the number of features (*Feat.*), the number of labels, label cardinality (*Card.*) and label density (*Dens.*). Moreover, we calculated the new measure of imbalanced data level of individual label, labelsets and labelsets-***p*** for each datasets. In detail, the imbalance level is raised by the LP transformation except for *Scene*, and the distribution of labelsets becomes less skewed by pruning. As shown in **Table 1**, the imbalance labelset ratio (*ImbalR*) of the labelsets can be as low as 16.6 (13.0 for labelsets-***p***), and the highest imbalance ratio happens to be 1903.3 (543.7 for labelsets-***p***). In this work, for each RankSVM method, the RBF kernel is chosen (the kernel parameter is set to the mean value of distances between each pair of two instances), and the regularization parameter $C$ is optimized by nested cross validation. We use macro-FM, micro-FM and ranking loss as evaluation criteria [6]. All the experiments are conducted by 10-fold cross-validation. The experimentation to our proposed CSRankSVM involves two stages:

***Experiment I****:* To exhibit the influences of imbalanced labelsets and the performance of our proposed approaches, the comparison is conducted between our two methods and the traditional RankSVM in order to validate the effectiveness of cost sensitive learning experimentally;

***Experiment II**:* We investigate the performance of CSRankSVM compared to six state-of-the-art methods. The results can confirm the advantages of our approach for multi-label data learning.

**Table 1.** Statistics for six benchmark datasets used in our experiments.

| **Datasets** | **Instances** | | **Statistics** | | | | ***ImbalR*** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Training** | **Test** | ***Feat.*** | ***Labels*** | ***Card.*** | ***Dens.*** | **Label** | Labelsets | **Labelsets-*p*** |
| Scene | 1211 | 1196 | 294 | 6 | 1.074 | 0.179 | 24.3 | 16.6 | 13.0 |
| Image | 1200 | 800 | 294 | 5 | 1.236 | 0.247 | 10.8 | 43.4 | 34.4 |
| Emotions | 391 | 202 | 72 | 6 | 1.868 | 0.311 | 25.7 | 69.3 | 48.2 |
| Plant | 558 | 390 | 440 | 12 | 1.079 | 0.090 | 55.2 | 381.1 | 132.4 |
| Human | 1864 | 1244 | 440 | 14 | 1.185 | 0.085 | 66.5 | 1278.1 | 384.0 |
| Yeast | 1500 | 917 | 198 | 14 | 4.237 | 0.303 | 36.7 | 1903.3 | 543.7 |

## 4.2 Experiment I

In this experiment, the comparison is conducted between our two methods and the basic RankSVM as well as **OS**RankSVM (RankSVM combined with random Over-Sampling) and **US**RankSVM (RankSVM combined with random Under-Sampling) [6] in terms of Macro-FM, Micro-FM, Ranking Loss and size of labelsets. Both re-sampling methods **OS**RankSVM and **US**RankSVM are among the most used prepro-cessing methods to equilibrate imbalanced datasets, and work in the labelset space as well. **Table 2** summarizes the performance of the compared algorithms.

**Table 2.** The performance of the three RankSVM methods

| Methods | Metric | Yeast | Human | Plant | Scene | Emotions | Image |
|---|---|---|---|---|---|---|---|
| RankSVM | Macro-FM | 0.3466 | 0.1637 | 0.1092 | 0.7084 | 0.6412 | 0.6358 |
| | Micro-FM | 0.6494 | 0.4306 | 0.3464 | 0.6981 | 0.6562 | 0.6310 |
| | Ranking | 0.1743 | 0.1370 | 0.1801 | 0.0745 | 0.1716 | 0.1508 |
| OSRankSVM | Macro-FM | 0.3669 | 0.1744 | 0.1680 | 0.7166 | 0.6494 | 0.6267 |
| | Micro-FM | 0.6579 | 0.4455 | 0.3704 | 0.7084 | 0.6615 | 0.6226 |
| | Ranking | 0.1728 | 0.1353 | 0.1725 | 0.0738 | 0.1749 | 0.1532 |
| USRankSVM | Macro-FM | 0.3369 | 0.1397 | 0.096 | 0.7137 | 0.6109 | 0.6176 |
| | Micro-FM | 0.6465 | 0.4095 | 0.3172 | 0.7046 | 0.6434 | 0.6094 |
| | Ranking | 0.1779 | 0.1471 | 0.1896 | 0.0735 | 0.1741 | 0.1582 |
| CSRankSVM | Macro-FM | 0.3930 | 0.2414 | 0.2468 | 0.7130 | 0.6719 | 0.6432 |
| | Micro-FM | 0.6626 | 0.4357 | 0.3595 | 0.7051 | 0.6735 | 0.6387 |
| | Ranking | 0.1653 | 0.1372 | 0.1656 | 0.0704 | 0.1525 | 0.1398 |
| CSRankSVM-p | Macro-FM | 0.3832 | 0.2334 | 0.2448 | 0.7170 | 0.6545 | 0.6654 |
| | Micro-FM | 0.6560 | 0.4330 | 0.3544 | 0.7094 | 0.6598 | 0.6591 |
| | Ranking | 0.1661 | 0.1360 | 0.1694 | 0.0683 | 0.1533 | 0.1351 |

The results shows that the phenomenon of imbalanced labelsets affects the multi-label datasets, and our methods with a cost sensitive learning strategy can improve the performance for different imbalance degrees. That is to say, the separating boundary is moved towards the majority class so that additional minority samples are correctly classified, but slightly more majority samples are misclassified. Moreover, although CSRankSVM-***p*** does not improve the performance of CSRankSVM on the most da-tasets, the preprocessing of pruning can reduce the number of labelsets, obtains an improvement in efficiency but with inevitably slight information loss.

What is most important is that even the nearly balanced dataset from the label level, such as *Scene* and *Image*, benefited from the proposed methods. Therefore, from the view of individual label, the proposed weighted CSRankSVM is applicable to not only the imbalanced multi-label datasets, but also the well balanced datasets so long as the imbalanced labelsets exist.

### 4.3 Experiment II

In this section, we experimentally compare our CSRankSVM with five existing multi-label classification approaches involving data transformation strategy and algorithm adaptation strategy. The competing methods are BP-MLL (Backpropagation for Multi-Label Learning) [12], BR (binary relevance) [13], HOMER(Hierarchy Of Multilabel classifiERs) [14], CC(Classifier Chain) [15], RakEL(RAndom *k* labELsets) [3],CLR (Calibrated label ranking)[16]. The results are given in Tables 3-5, in which the best performance for each dataset is highlighted. The numbers in parentheses denote the rank of the algorithm among the compared algorithms.

**Table 3.** The comparison of the proposed methods with respect to Macro-FM

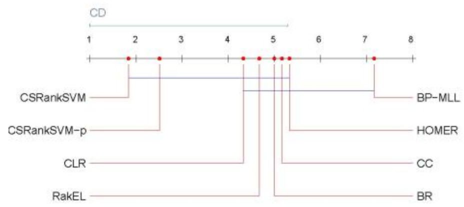
| Methods | Yeast | Human | Plant | Scene | Emotions | Image | AvgRank |
|---|---|---|---|---|---|---|---|
| BP-MLL | 0.3392(8) | 0.0013(8) | 0.0219(8) | 0.0537(8) | 0.6379(3) | 0.2917(8) | 7.17 |
| BR | 0.3920(5) | 0.1527(3) | 0.1387(5) | 0.6285(5) | 0.5891(6) | 0.5414(6) | 5.00 |
| HOMER | **0.4066(1)** | 0.1430(5) | 0.1549(4) | 0.6027(7) | 0.5601(8) | 0.5348(7) | 5.33 |
| CC | 0.3966(3) | 0.1498(4) | 0.1121(6) | 0.6126(6) | 0.5732(7) | 0.5437(5) | 5.17 |
| RakEL | 0.4031(2) | 0.1304(7) | 0.0765(7) | 0.7070(3) | 0.6313(4) | 0.6355(3) | 4.33 |
| CLR | 0.3834(6) | 0.1314(6) | 0.1552(3) | 0.6442(4) | 0.6242(5) | 0.5568(4) | 4.67 |
| CSRankSVM | 0.3930(4) | **0.2414(1)** | **0.2468(1)** | 0.7130(2) | **0.6719(1)** | 0.6432(2) | **1.83** |
| CSRankSVM-p | 0.3832(7) | 0.2334(2) | 0.2448(2) | **0.7170(1)** | 0.6545(2) | **0.6654(1)** | 2.50 |

**Table 4.** The comparison of the proposed methods with respect to Micro-FM

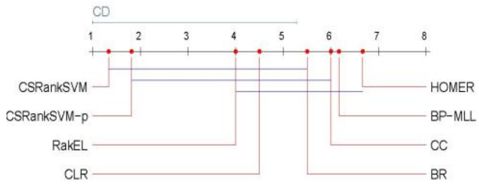
| Methods | Yeast | Human | Plant | Scene | Emotions | Image | AvgRank |
|---|---|---|---|---|---|---|---|
| BP-MLL | 0.6452(3) | 0.0027(8) | 0.0234(8) | 0.1670(8) | 0.6608(2) | 0.3583(8) | 6.17 |
| BR | 0.5857(7) | 0.2949(5) | 0.2376(4) | 0.6194(5) | 0.6055(6) | 0.5395(6) | 5.50 |
| HOMER | 0.5858(6) | 0.2725(7) | 0.2362(5) | 0.5927(7) | 0.5748(8) | 0.5337(7) | 6.67 |
| CC | 0.5499(8) | 0.2960(4) | 0.2078(6) | 0.6001(6) | 0.5868(7) | 0.5419(5) | 5.83 |
| RakEL | 0.6254(4) | 0.2905(6) | 0.1267(7) | 0.6977(3) | 0.6467(4) | 0.6318(3) | 4.50 |
| CLR | 0.6158(5) | 0.3023(3) | 0.2564(3) | 0.6276(4) | 0.6364(5) | 0.5545(4) | 4.00 |
| CSRankSVM | **0.6626(1)** | **0.4357(1)** | **0.3595(1)** | 0.7051(2) | **0.6735(1)** | 0.6387(2) | **1.33** |
| CSRankSVM-p | 0.6560(2) | 0.4330(2) | 0.3544(2) | **0.7094(1)** | 0.6598(3) | **0.6591(1)** | 1.83 |

**Table 5.** The comparison of the proposed methods with respect to Ranking Loss

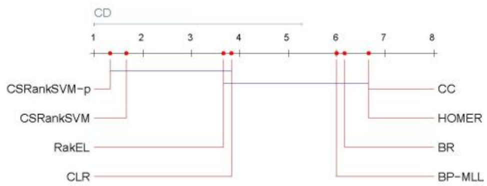
| Methods | Yeast | Human | Plant | Scene | Emotions | Image | AvgRan |
|---|---|---|---|---|---|---|---|
| BP-MLL | 0.1748(3) | 0.3934(5) | 0.4876(8) | 0.3986(8) | 0.1825(4) | 0.3430(8) | 6.00 |
| BR | 0.3097(6) | 0.4163(8) | 0.4772(6) | 0.2465(6) | 0.2941(6) | 0.2968(5) | 6.17 |
| HOMER | 0.3287(8) | 0.3989(7) | 0.4362(5) | 0.2345(5) | 0.3091(8) | 0.3055(7) | 6.67 |
| CC | 0.3238(7) | 0.3935(6) | 0.4861(7) | 0.2489(7) | 0.3028(7) | 0.2995(6) | 6.67 |
| RakEL | 0.2135(5) | 0.2311(4) | 0.1961(3) | 0.0998(3) | 0.1872(5) | 0.1807(3) | 3.83 |
| CLR | 0.1783(4) | 0.1571(3) | 0.2085(4) | 0.1011(4) | 0.1699(3) | 0.1917(4) | 3.67 |
| CSRankSVM | **0.1653(1)** | 0.1372(2) | 0.1656(2) | 0.0704(2) | **0.1525(1)** | 0.1398(2) | 1.67 |
| CSRankSVM-p | 0.1661(2) | **0.1360(1)** | **0.1694(1)** | **0.0683(1)** | 0.1533(2) | **0.1351(1)** | **1.33** |




(a) Significant test of Macro-FM




(b) Significant test of Micro-FM




(c) Significant test of Ranking Loss

**Fig. 4.** Significant test of comparable methods

To better understand the results of our techniques when compared to the other classification approaches, we performed a statistical analysis of our results. Firstly, a non-parametric Friedman test is used to determine that there is a statistically significant difference between the rankings of the classifiers in terms of G-mean and AUC. Consequently, we reject the null-hypothesis stating that all algorithms perform equally in mean ranking. Based on this rejection, the Nemenyi post-hoc test is used to compare all classifier to each other. It can be found from **Fig. 4** that our methods statistically outperform BP-MLL in terms of Macro-FM, statistically outperform BP-MLL and HOMER in terms of Micro-FM, and statistically outperform CC, BR, BP-MLL and HOMER in terms of ranking loss.

# 5 Conclusion

This paper studies the challenges posed by the labelset imbalance problem. We introduce a new metric aimed to measure the imbalance level in multi-label datasets, and two cost sensitive methods designed to reduce the imbalance level of labelsets multi-label datasets are proposed. The experimental results on some benchmark multi-label data have demonstrated that the proposed methods provide a very competitive solution to other existing state-of-the-arts multi-label data classification methods.

## 6 Acknowledgments

This research was supported by the National Natural Science Foundation of China (61502091), the Fundamental Research Funds for the Central Universities (N140403004), and the Postdoctoral Science Foundation of China (2015M570254).


## Reference


1. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering **26**(8), 1819-1837 (2014)
2. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining **3**(3), 1-13 (2007)
3. STsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering **23**(7), 1079-1089 (2011)
4. Lo, H. Y., Lin, S. D., Wang, H. M.. Generalized k-labelsets ensemble for multi-label and cost-sensitive classification. IEEE Transactions on Knowledge and Data Engineering **26**(7), 1679-1691 (2014)
5. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263-1284 (2009)
6. Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3-16 (2015)
7. Tahir, M.A., Kittler, J., Yan, F.: Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recognition **45**,3738-3750 (2012)
8. Elisseeff, A., Weston, J.: Kernel Methods for Multi-labelled Classification and Categorical Regression Problems. Technical Report, BIOwulf Technologies http://www.kyb.tuebingen.mpg.de/bs/people/weston/publications (2001)
9. Bao, L., Juan, C., Li, J., & Zhang, Y.: Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing**172**, 198-206 (2016)
10. Cao, P., Zhao, D., & Zaiane, O.: An optimized cost-sensitive SVM for imbalanced data learning. In: Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013), pp. 1558–1564 (2013)
11. Read, J.: A pruned problem transformation method for multi-label classification. In: Proceedings of the 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), **143150** (2008)
12. Zhang, M.L., Zhou, Z.H.: Multi-label Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering **18**(10), 1338-1351 (2006)
13. Boutell, M.R. Luo, J., Shen, X., Brown,C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9), 1757-1771 (2004)
14. Tsoumakas, G., Katakis, I., & Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: I Proceedings of ECML/PKDD Workshop on Mining Multidimensional Data (MMD'08), pp. 30-44 (2008)
15. Reed, J., Pfahringer, B., Holmes, G.: Classifier chain for multi-label classification. Machine Learning **85** (3), 333-359 (2011)
16. Fürnkranz, J., Hüllermeier, E., Mencía, E. L., Brinker, K.: Multilabel classification via calibrated label ranking. Machine learning **73**(2), 133-153 (2008)