CP-RA*k*EL: Improving Random *k*-labelsets with Conformal Prediction for Multi-label Classification

Fan Yang YANG@XMU.EDU.CN and **Xiaolu Gan** XLGAN@STU.XMU.EDU.CN
Department of Automation
Xiamen University
Xiamen,361005,P.R.China

Huazhen Wang WANGHUAZHEN@HQU.EDU.CN
College of Computer Science and Technology
Huaqiao University
Xiamen,361021, P.R.China

Lei Feng FENGLEI.SZ@CCB.COM
Shenzhen Branch
China Construction Bank
Shenzhen,518000,P.R.China

Yongxuan Lai LAIYX@XMU.EDU.CN
Software School
Xiamen University
Xiamen, 361005,P.R.China

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Multi-label conformal prediction has attracted much attention in the conformal predictor (CP) community. In this article, we propose to combine CP with random k-labelsets (RAkEL) method, which is state-of-the-art multi-label classification method for large number of labels. In the framework of RAkEL, the original problem is reduced to a number of small-sized multi-label classification tasks by randomly breaking the initial set of labels into a number of small-sized labelsets, and then label powerset (LP) method is employed on these tasks respectively. In this work, ICP-RF, an inductive conformal predictor based on random forest, is used in each multi-label task in order to get p-values for predictions of the LP model, and then the predictions are aggregated to get a final result. Experimental results on six benchmark datasets empirically demonstrate the calibration property of ICP-RF as LP models, and show that conformal prediction can significantly improve the performances of the proposed approach, which is called RAkEL. However, the validity property of CP does not hold in CP-RAkEL. In the future work we will study how to use some new CP techniques to calibrate the new method.

Keywords: multi-label classification, conformal prediction, random k-labelsets

1. Introduction

Multi-Label Learning (MLL) has been widely used in many real-world tasks, such as text categorization, image retrieval, and bioinformatics (Zhang and Zhou, 2014). The specific property of Multi-Label Learning, i.e. the instance always can be assigned to more than one label and the task is to predict a set of labels, is similar to the learning paradigm of conformal prediction, which serves as a classifier for region prediction. Consequently, recent studies focused on conducting multi-label learning with conformal predictors and proposing a new framework of multi-label conformal prediction (Wang et al., 2014; Papadopoulos, 2014; Wang et al., 2015).

In multi-label conformal prediction, researchers concern the concept of calibration property in multi-label learning, which refers to the error rate of prediction, i.e., the probability of falsely predicting the labels being bounded by a threshold function. Thus, a well-calibrated MLL classifier highlights providing reliable prediction, for example, a prediction at confidence level 90% indicates at least 90% certainty about the predicted labelset including the true multi-labels. In a previous study, we first adapted conformal predictor to produce calibrated MLL prediction (Wang et al., 2014). A multi-label example is transformed into several single-label examples, and then a conformal predictor is applied to output the prediction labelset. Meanwhile, Papadopoulos proposed another method using conformal predictor for MLL (Papadopoulos, 2014) which maps each labelset as a new single-label to fit the framework of conformal predictor. Wang et al. proposed a binary relevance method to use conformal predictor for multi-label learning and conducted a comparison with the other two methods (Wang et al., 2015). Recently Wang et al. proposed a conceptually novel framework of Multi-Label Conformal Prediction (Wang et al., 2017), which associates the multi-label prediction with a reliable measure of confidence and the implementation of multi-label conformal prediction is a combination of Pattern Transformation (PT) and Conformal Prediction (CP) (Shafer and Vovk, 2008), three effective PT methods are applied to respectively establish three practical implementation models. Among them are Power Set Multi-Label Conformal Predictor (PS-MLCP) which combines Label Powerset (LP) and CP, Binary Relevance Multi-Label Conformal Predictor (BR-MLCP) which combines of Binary Relevance method and CP, Instance Reproduction Multi-Label Conformal Predictor (IR-MLCP) which combines of Instance Reproduction method and CP.

In this article we focuses on the label powerset (LP) Multi-Label Conformal Predictor (PS-MLCP), which combines LP method and conformal predictor. LP has the advantage of taking label correlations into consideration and achieves better performance compared to computationally faster methods like binary relevance (BR). However, it is difficult for LP to deal with real applications with large number of labels, which may generate large number of labelsets in the training set. The large number of these labelsets not only raises the computational cost of LP on one hand, but always makes it a highly imbalanced classification task as many of these labelsets are usually associated with very few training instances (Tsoumakas et al., 2011a). Moreover, LP can only predict labelsets observed in the training set which limits its use for unseen new labelsets (Tsoumakas et al., 2011a; Tsoumakas and Vlahavas, 2007). Noted that, similar to its implemented LP classifiers, PS-MLCP also faces with aforementioned issues rising from large number of labels.

Motivated by the idea of RAkEL (RAndom k labELsets) method, which proposes randomly breaking the initial set of labels into a number of small-sized labelsets, and employing LP to train a corresponding multi-label classifier, in this work, we proposed a novel approach called CP-RAkEL which attempt to combine RAkEL with conformal prediction in order to deal with large number of labels. Specifically, we implement our previous CP-RF method (Yang et al., 2009) in the framework of RAkEL to construct LP classifiers, and we named it ICP-RF as it is conducted in the inductive manner (Papadopoulos, 2008). For the hard prediction of each test instance we combine the predicted labelsets associated with the largest p-values generated from each ICP-RF model by average votes. For the region prediction, for each test instances we average on the p-values generated from different ICP-RF models to get a final p-value for each label and then test the calibration property of the proposed method. Experimental results on six benchmark datasets empirically demonstrate the calibration property of ICP-RF models as label powerset (LP) classifiers, and show that conformal prediction can significantly improve the performances of RAkELo method.

The organization of this article is as follows. In the next section the related work of random k-labelsets for multi-label classification and multi-label conformal prediction are introduced. In Section 3, the implementation model of CP-RAkEL is proposed. And in Section 4, the classification accuracy and calibration performance of the new method are empirically illustrated. Finally, the conclusion is presented in Section 5.

2. Related work

2.1. Random k-labelsets for multi-label classification

RAkEL aims to help LP to deal with large number of label sets (Tsoumakas et al., 2011a). In order to deal with the aforementioned problems of LP, Tsoumakas et al. (2011a) propose to randomly break the initial set of labels into a number of small-sized labelsets, and hence get a number of small-sized, computationally simpler and less skewed multi-label classification tasks and then employ LP on these tasks respectively. The parameter k specifies the size of the labelsets and the complexity of the multi-label classification tasks. There are two versions of RAkEL.Given a specific value of k, RAkELd partitions the whole label set L with size M randomly into $N = [M/k]$ disjoint labelsets R_j, $j = 1, ..., N$, and partitions the training set into subsets D_j correspondingly, and then RAkELd learns m multi-label classifiers h_j using LP. For RAkELo, let the term L^k denote the set of all distinct k-labelsets of L. The size of L^k is given by the binomial coefficient: $L^k = \binom{M}{k}$. Given a size of labelsets k and a number of desired classifiers $m \leq |L^k|$, RAkELo initially selects m k-labelsets R_i, $i = 1, ..., m$ from the set L^k via random sampling without replacement. Note that in this case the labelsets may overlap. The computational complexity of RAkEL is linear with respect to the number of LP classifiers and their complexity. Further, the number of LP classifiers is linear with respect to M in the case of disjoint labelsets while in the case of overlapping labelsets a value of m that is linear with respect to M is also able to achieve good performances empirically.

Given a new instance x, the predictions $h_i(x, j)$ are gathered in order to build the final multi-label classification outputs. In the case of RAkELo each model h_i provides binary predictions $h_i(x, j)$ for each label j in the corresponding k-labelset R_i and then employ the average votes rule for the fusion of LP classifier outputs.

2.2. Multi-label conformal prediction

Multi-Label Conformal Prediction (MLCP) (Wang et al., 2017) aims to not only provide prediction labelset but indicate a reliable measure of confidence for the prediction. Given a multi-label training dataset $Z(n) = (Z_1, Z_2, ..., Z_n)$ and a test instance x_t, where each training example $Z_i = (x_i, Y_i)$, $X \in R^d$ and Y_i is a subset of whole labels set. Given a predefined significance level ε, MLCP predicts the test instance x_t with a prediction which is associated with a measure of confidence $1 - \varepsilon$. In a nutshell, the framework of multi-label conformal prediction can be divided into three steps: pattern transformation, algorithmic randomness test, labelset adaption. Firstly, pattern transformation technique is applied to transform the original multi-label training dataset into single-label training dataset. Secondly, the basic procedure of algorithmic randomness test is conducted to produce p-value measuring the significance of the test. Thirdly, the output p-values are applied to construct the prediction for the test instance.

2.3. The CP-RF method

In (Yang et al., 2009) we propose a new algorithm called CP-RF which hedges the predictions of random forest with conformal predictor. Random forest classifier naturally leads to a dissimilarity measure between examples in a "strange" space rather than a Euclidean measure. After a RF is grown, since an individual tree is unpruned, the terminal nodes will contain only a small number of observations. Given a random forest of size *ntree*: $f = T_1, ..., T_{ntree}$ and two examples x_i and x_j, propagate the examples down all the trees within f. Let $D_i = T_1, ..., T_{ntree}$ be tree nodes for x_i and x_j,on all the trees respectively, a random forest proximity between the two examples is defined as,

$$\text{prox}(i, j) = \frac{1}{\text{ntree}} \sum_{t=1}^{\text{ntree}} I(T_{t,i}, T_{t,j}) \tag{1}$$

where $I(T_{t,i}, T_{t,j}) = \begin{cases} 1 & \text{if } T_{t,i} = T_{t,j} \\ 0 & \text{otherwise} \end{cases}$.

If instance i and j both land in the same terminal node,the proximity between them is increased by one, this forms a matrix $(prox(i,j))_{N \times N}$, which is symmetric, positive definite and bounded above by 1, with the diagonal elements equal to 1, and N is the total number of cases (Breiman, 2001). Then in the distance matrix of the N cases defined by the proximity matrix, we designed nonconformity scores using outlier measure of random forest (Yang et al., 2009).

3. Improving Random k-labelsets with Conformal Prediction for large number of labels

In this section, we propose to combine RAkEL method with conformal prediction. The motivation is two folds: first, as a state-of-the-art approach, RAkEL is able to improve the performance of label powerset (LP) method when faced with large number of labels. Combining RAkEL and conformal prediction could not only make multi-label conformal prediction with LP (PS-MLCP) suitable for large number of labels, but improve the performance of RAkEL. Second, the effectiveness and calibration of CP-RAkEL in multi-label classification would be investigated.

In this paper we mainly focus on RAkELo which usually performs better than RAkELd. Looking into the classification mechanism of RAkELo, for a test point, each LP classifier outputs a hard prediction which associates with a value of 0/1 for the predicted labels and then the ensemble averages over the predictions and get a final prediction with average voting rule. In this manner the confidence of each LP classifier on that test points is neglected while in this work we propose a new method called CP-RAkEL which introduces the confidence of each LP predictions into the voting mechanism and improves the performance of RAkELo. Moreover, the final outputs of CP-RAkEL can also be labelsets associated with p-values for each label and hence the confidence of predictions can be quantified.

In the framework of RAkELo, we implement Inductive CP-RF (ICP-RF) as the base LP classifier. After a random forest model is built, we get proximity matrix using the same approach in Section 2.3. Different from the nonconformity measures calculated by outliers used in (Yang et al., 2009), we conducted CP-kNN in the distance matrix defined by proximity, i.e. designing nonconformity scores as follows,

$$a_i = \frac{\sum_{j=1}^{K} D_{ij}^{yi}}{\sum_{j=1}^{K} D_{ij}^{-y_i}} \tag{2}$$

where D_{ij}^{yi} is the jth shortest distance between instance x_i and the instances with the same label as x_i, and D_{ij}^{-yi} is the jth shortest distance between instance $x\ i$ and the instances with different label, and K is the number of nearest neighbors. For each training instance, its nonconformity score can be calculated offline according to the trained forest model. When a test point comes, we put it into the forest model and get its proximities to all the training instances and then also calculate its nonconformity score with Eq(2), which is conducted in an inductive manner. The pseudocode of ICP-RF is shown in Algorithm 1 and the training process of CP-RAkEL is shown in Algorithm 2 respectively.

For the hard prediction of a multi-label test point, we first output the labelset associated with the largest p-value in each LP model. Take an example in Table 1, the size of the whole label set M =7. And we set parameter k =3, that is, for each multi-label classification task, three labels are randomly selected. Then we randomly get nine tasks from the original problem and learn a LP model for each task. For model h_1, the label set is $\{l_1, l_3, l_7\}$, and the predicted labelset with the largest p-value 0.95 is $\{l_1, l_3\}$, so the output are labels l_1 and l_3 associated with p-value 0.95 while the p-value of l_7 is set to be zero. Then we get final prediction by average votes as shown in Table 2. And the detailed prediction process is shown in Algorithm 3.

To test the calibration property, for reliable prediction CP-RAkEL outputs average p-values for each label over the outputs of all LP models. For simplicity, only the labelset associated with the largest p-value in each LP is considered. For a predefined significance level ϵ, the region prediction is given according to the results of CP-RAkEL, which is shown in Algorithm 4.

4. Experimental results

This section provides empirical results and analysis of the experiments. Specifically, Section 4.1 describes the datasets and Section 4.2 the evaluation measures for multi-label

```
Algorithm 1: Inductive CP-RF
Input:  training set Z(n)=$\{Z_1, Z_2, ..., Z_n\}$,Label set $Y$, testing instances
        X(t)=$\{X_1, X_2, ..., X_t\}$
Output: $p$-values matrix P.

Train random forest model on $Z(n)$;
Generate proximity matrix $Prox$ for $Z(n)$;
$D \leftarrow 1 - Prox$ ;
for $i \leftarrow 1$ to $n$ do
    $distance1 \leftarrow \sum_{j=1}^{K} D_{ij}^{yi}$;
    $distance2 \leftarrow \sum_{j=1}^{K} D_{ij}^{-yi}$;
    $a_i \leftarrow distance1/distance2$;
end
Generate proximity matrix $Prox'$ for both $Z(n)$ and$X(t)$;
$D' \leftarrow 1 - Prox'$ ;
for $i \leftarrow 1$ to $t$ do
    for $y \in Y$ do
        $distance3 \leftarrow \sum_{j=1}^{K} D'^{yi}_{ij}$;
        $distance4 \leftarrow \sum_{j=1}^{K} D'^{-yi}_{ij}$;
        $a_{iy} \leftarrow distance3/distance4$;
        $count \leftarrow |a_1, a_2 ..., a_n > a_{iy}|$;
        $count1 \leftarrow |a_1, a_2 ..., a_n = a_{iy}|$;
        $p_{iy} \leftarrow (count + \tau * count1)/(m + 1)$($\tau$ is a random value in [0,1]);
    end
end
```

```
Algorithm 2: Training Process of CP-RAkEL
Input:  Set of labels $L$ of size $M$,training set $Z(n)$,labelset size $k$,number of models
        $N \leq \binom{M}{k}$.
Output: ICP-RF classifiers $h_i, i = 1, ..., N$

$S \leftarrow L^k$;
for $i \leftarrow 1$ to $N$ do
    $R_i \leftarrow$ a k-labelset randomly selected from S;
    train an ICP-RF classifier $h_i$ based on $Z(n)$and $R_i$;
    $S \leftarrow S \setminus \{R_i\}$;
end
```

Table 1: An example of the output p-values of CP-RAkEL (k=3,N=9,M=7)

model	**labelset**	l_1	l_2	l_3	l_4	l_5	l_6	l_7
h1	l_1, l_3, l_7	0.95	–	0.95	–	–	–	0
h2	l_1, l_4, l_6	0.86	–	–	0.86	–	0.86	–
h3	l_1, l_2, l_7	0	0.93	–	–	–	–	0
h4	l_2, l_4, l_6	–	0.89	–	0	–	0.89	–
h5	l_3, l_5, l_6	–	–	0.98	–	0	0	–
h6	l_2, l_6, l_7	–	0	–	–	–	0	0.91
h7	l_3, l_5, l_6	–	–	0	–	0.69	0	–
h8	l_1, l_2, l_4	0.79	0	–	0	–	–	–
h9	l_2, l_3, l_5	–	0	0.94	–	0.94	–	–

Table 2: An example of the classification process of RAkEL(k=3,N=9,M=7)

model	**labelset**	l_1	l_2	l_3	l_4	l_5	l_6	l_7
h1	l_1, l_3, l_7	1	–	1	–	–	–	0
h2	l_1, l_4, l_6	1	–	–	1	–	1	–
h3	l_1, l_2, l_7	0	1	–	–	–	–	0
h4	l_2, l_4, l_6	–	1	–	0	–	1	–
h5	l_3, l_5, l_6	–	–	1	–	0	0	–
h6	l_2, l_6, l_7	–	0	–	–	–	0	1
h7	l_3, l_5, l_6	–	–	0	–	1	0	–
h8	l_1, l_2, l_4	1	0	–	0	–	–	–
h9	l_2, l_3, l_5	–	0	1	–	1	–	–
average	votes	3/4	2/5	3/4	1/3	2/3	2/5	1/3
predicted	labels	1	0	1	0	1	0	0

Algorithm 3: Hard Prediction with CP-RAkEL

```
Input: Set of labels $L$ of size $M$, number of models $N$, $k$-labelsets $R_i$, corresponding
       ICP-RF classifier $h_i$, a test instance $x$
Output: $Result$
for $j \leftarrow 1$ to $M$ do
    $Sum_j \leftarrow 0$;
    $Votes_j \leftarrow 0$;
end
for $i \leftarrow 1$ to $N$ do
    $l_{max} \leftarrow$ the labelset which has the largest $p$-value for instance $x$ in $h_i$;
    $y_{max} \leftarrow$ Mapping $l_{max}$ class into $L$;
    forall $j \in R_i$ do
        if $j \in y_{max}$ then
            $Sum_j \leftarrow Sum_j + 1$;
        end
        else
            $Sum_j \leftarrow Sum_j$;
        end
        $Votes_j \leftarrow Votes_j + 1$;
    end
end
for $j \leftarrow 1$ to $M$ do
    $Avg_j \leftarrow Sum_j / Votes_j$;
    if $Avg_j > 0.5$ then
        $Result_j \leftarrow 1$;
    end
    else
        $Result_j \leftarrow 0$;
    end
end
```

Algorithm 4: Reliable Prediction with CP-RAkEL

```
Input: Set of labels $L$ of size $M$, number of models $N$, $k$-labelsets $R_i$, corresponding
       ICP-RF classifier $h_i$, a test instance $x$, a predefined significance level $\epsilon$
Output: $Result$
for $j \leftarrow 1$ to $M$ do
    $Sum_j \leftarrow 0$;
    $Votes_j \leftarrow 0$;
end
for $i \leftarrow 1$ to $N$ do
    $l_{max} \leftarrow$ the labelset which has the largest $p$-value for instance $x$ in $h_i$;
    $y_{max} \leftarrow$ Mapping $l_{max}$ class into $L$;
    forall $j \in R_i$ do
        if $j \in y_{max}$ then
            $Sum_j \leftarrow Sum_j + h_i(x, l_{max})$;
        end
        else
            $Sum_j \leftarrow Sum_j$;
        end
        $Votes_j \leftarrow Votes_j + 1$;
    end
end
for $j \leftarrow 1$ to $M$ do
    $Avg_j \leftarrow Sum_j / Votes_j$;
    if $Avg_j > \epsilon$ then
        $Result_j \leftarrow 1$;
    end
    else
        $Result_j \leftarrow 0$;
    end
end
```

classification. In Section 4.3 we give an example to show the calibration property of ICP-RF as a specific LP classifier. In Section 4.4 we show the results of hard predictions of our method in comparison with RAkEL. In Section 4.5 the calibration property of CP-RAkEL is demonstrated on two datasets.

4.1. Description of the datasets

The datasets used in this experiment are all standard datasets in Mulan (Tsoumakas et al., 2011b). In the experiments, six datasets in different applications are tested and the characteristics of the datasets are shown in the following table. In Table 3, *instances* represent the number of instances, *nomial* and *numeric* indicates the number of categorical attributes and numeric attributes of the dataset respectively, *labels* represents the number of labels, and *labelsets* indicates the number of distinct labelsets.

Table 3: Descriptions of the datasets used in experiments

name	**domain**	**instances**	**nominal**	**numeric**	**labels**	**cardinality**	**labelsets**
emotions	music	593	0	72	6	1.869	27
scene	images	2407	0	294	6	1.074	15
yeast	biology	2417	0	103	14	4.237	198
Birds	audio	645	2	258	19	1.014	133
flags	images	194	9	10	7	3.392	54
CAL500	music	502	0	68	174	26.044	502

4.2. Evaluation Measures for Multi-label Classfication

For the comparison with RAkELo, three popular metrics for multi-label classification are used to evaluate the performance of hard predictions of our method.

Jaccard similarity score The Jaccard similarity coefficient (Zhang and Zhou, 2014) of the i-th example, with a ground truth labelset and predicted labelset, is defined as

$$J(Y_i, Z_i) = \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}. \tag{3}$$

Zero-one loss For each example, the entire set of labels must be correctly predicted; otherwise the loss for that sample is equal to one (Zhang and Zhou, 2014).

$$\text{zero\_one\_loss} = \frac{1}{N}\sum_{1}^{N} L_{0-1}(Y_i, Z_i) \tag{4}$$

where

$$L_{0-1}(Y_i, Z_i) = 1(Y_i \neq Z_i) \tag{5}$$

F_1 measure The F_1 measure is the harmonic mean of precision and recall. Consider a binary classification. Given the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), F_1 measure is defined as follows

$$F_1(TP, FP, FN) = \frac{2 * TP}{2 * TP + FP + FN} \tag{6}$$

In multi-label classification evaluation, macro-averaged and micro-averaged of F_1 measure are widely used, and in this paper we show the results with respect to micro F_1. Let $TP_\lambda, FP_\lambda, TN_\lambda$ and FN_λ be the number of true positives, false positives, true negatives and false negatives after binary evaluation for a label λ in the label set with size M, then the micro-averaged F_1 measure is calculated as follows (Tsoumakas et al., 2011a),

$$F_{1\text{ micro}} = F_1\left(\sum_{\lambda=1}^{M} TP_\lambda, \sum_{\lambda=1}^{M} FP_\lambda, \sum_{\lambda=1}^{M} FN_\lambda\right) \tag{7}$$

4.3. The calibration of ICP-RF as LP model

To show the calibration property of ICP-RF, we tested the calibration of the algorithm on randomly-partitioned LP problems. Figure 1 shows results on *emotions* and *birds* datasets. The results take the averages on 10-fold cross-validation. It can be seen from Figure 1 that the error rate fluctuates near the diagonal baseline, indicating that the error rate is calibrated. It also indicates that the p-values output by the LP classifiers are accurate and reliable which can be used to improve RAkEL. The performances of other LP classifiers on all the datasets are similar.

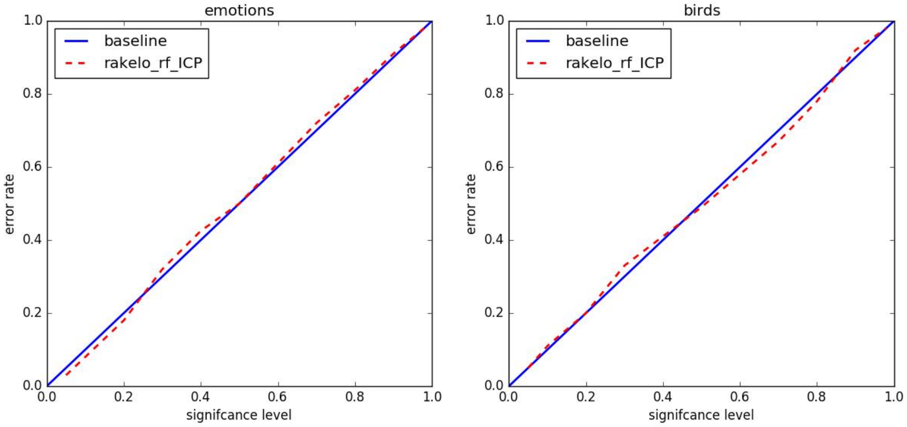

Figure 1: The calibration property of ICP-RF as LP classifiers

4.4. Comparison on hard predictions

In this section we compare our method with RAkEL where we use random forest as base classifier. The experiments were conducted on standard multi-label datasets in 10-fold cross-validation.The parameters of ICP-RF and random forest are set to be default values. The parameters that are adjusted in the experiments are k (the number of randomly selected labels), K (in ICP-RF, kNN is used to calculate the nonconformity scores) and the number of LP models for RAkELo. Through experiments, we found that there are some potential relationships between the parameters. When we set $k = 2$, each LP problem is a multi-class classification task with about four classes on average, in this case setting K = 9 achieves best results; when $k = 3$, each LP problem is a multi-class classification task with about eight classes on average, and in this case setting K = 16 gets best results. Empirically, it seems that K should take 2 times the number of classes. The number of LP models N is between $2M$ and $3M$ (M is the number of labels). In the following, the results of three evaluation metrics under parameter settings $k = 2$, K= 9, $N = 3M$ are shown in Table 4. It shows that in terms of Micro F_1 and Jaccard similarity, the proposed approach outperforms the popular RAkEL with random forest as base LP classifiers on all datasets except CAL500. While with respect to zero-one loss, these two approaches have comparable performances.

Table 4: Comparison of prediction accuracy on six datasets

	RAkELo with RF			RAkELo with ICP-RF		
	micro F1	**jaccard**	**zero-one-loss**	**micro F1**	**jaccard**	**zero-one-loss**
emotions	67.46 ± 0.17%	54.73 ± 0.16%	66.77 ± 0.17%	69.66 ± 0.14%	59.26 ± 0.18%	65.61 ± 0.23%
scene	62.83 ± 2.89%	50.21 ± 2.67%	52.12 ± 2.92%	67.64 ± 2.40%	59.47 ± 2.78%	50.16 ± 3.02%
yeast	63.26 ± 0.01%	50.96 ± 0.01%	81.42 ± 0.04%	64.04 ± 0.05%	51.31 ± 0.10%	87.04 ± 0.59%
Birds	44.66 ± 0.09%	60.31 ± 0.31%	46.64 ± 0.48%	44.89 ± 0.10%	61.67 ± 0.27%	45.89 ± 0.43%
flags	74.84 ± 0.39%	61.20 ± 0.75%	77.24 ± 2.28%	75.89 ± 0.32%	62.39 ± 0.58%	77.13 ± 1.09%
CAL500	32.26 ± 0.06%	19.73 ± 0.03%	100.00 ± 0.0%	29.36 ± 0.02%	18.78 ± 0.36%	100.00 ± 0.0%

4.5. Test of the calibration property of CP-RAkEL

For the multi-label conformal prediction using CP-RAkEL, we record the error rates under different significant levels. Selected results of CP-RAkEL on *birds* and *emotions* datasets are shown in Figure 2 which show that the calibration property of CP does not hold in CP-RAkEL on *emotions* data. The reason mainly lies in the generation of p-values for the final results. First, only the labelsets with the largest p-value from each LP model are considered. Second, the p-values are averaged over the results of all LP models. In the future work we will study how to use some new CP techniques to calibrate CP-RAkEL.

5. Conclusions

In this article we combine the Random k-Labelsets method with ICP-RF in order to improve the performances of RAkEL on multi-label classification tasks with large number of labels. Experiment results demonstrate the effectiveness of our approaches. However, as illustrated in the experimental results, the validity property of CP does not hold in CP-Random k-Labelsets, which mainly attribute to the calculation of the p-values for the final results. Only the labelsets with the largest p-value from each model are considered and then averaged. In the future, we will explore more efficient and effective multi-label conformal prediction method for large number of labels to face with the need of real applications.

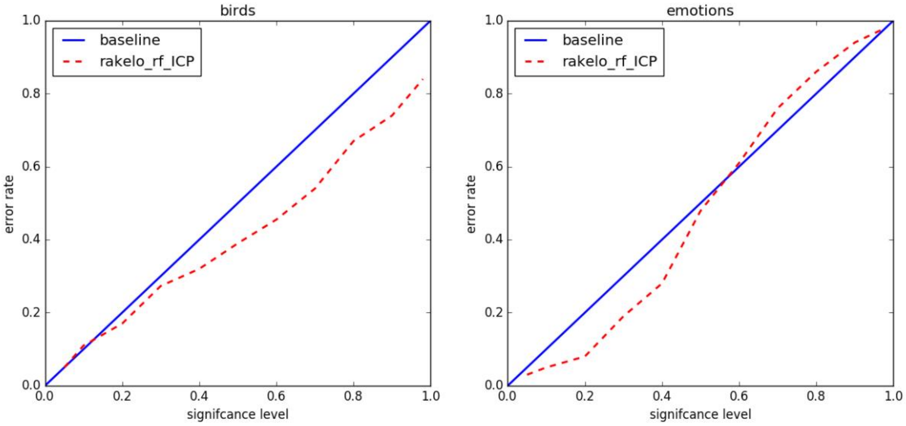

Figure 2: Test of the calibration property of CP-RAkEL

Acknowledgments

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. Fan Yang also thanks Prof. Samuel Kaski from the Department of Computer Science, Aalto University for hosting him as a visiting scholar in Aalto University. This work is supported by the State Scholarship Fund of China Scholarship Council under Grant No.201506315008 and Natural Science Foundation of China under Grant No.61673324 and No.61672441, and the Fundamental Research Funds for the Central Universities of China under Grant No.20720160085.

References

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher, 2008.

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 241–250. Springer, 2014.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *European Conference on Machine Learning*, pages 406–417. Springer, 2007.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 1079–1089, 2011a.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(Jul):2411–2414, 2011b.

Huazhen Wang, Xin Liu, Bing Lv, Fan Yang, and Yanzhu Hong. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine. *PloS one*, 9(6):e99565, 2014.

Huazhen Wang, Xin Liu, Ilia Nouretdinov, and Zhiyuan Luo. A comparison of three implementations of multi-label conformal prediction. In *International Symposium on Statistical Learning and Data Sciences*, pages 241–250. Springer, 2015.

Huazhen Wang, Cheng Wang, Ilia Nouretdinov, and Zhiyuan Luo. The framework and three implementation models of multi-label conformal prediction for theoretically reliable prediction. *submitted*, 2017.

Fan Yang, Hua-zhen Wang, Hong Mi, Wei-wen Cai, et al. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10 (1):S22, 2009.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.