A Cross-Conformal Predictor for Multi-label Classification

Harris Papadopoulos

HAL Id: hal-01391051
https://hal.inria.fr/hal-01391051

Submitted on 2 Nov 2016

A Cross-Conformal Predictor for Multi-label Classification

Harris Papadopoulos

Computer Science and Engineering Department, Frederick University,
7 Y. Frederickou St., Palouriotisa, Nicosia 1036, Cyprus
h.papadopoulos@frederick.ac.cy

Abstract. Unlike the typical classification setting where each instance is associated with a single class, in multi-label learning each instance is associated with multiple classes simultaneously. Therefore the learning task in this setting is to predict the subset of classes to which each instance belongs. This work examines the application of a recently developed framework called Conformal Prediction (CP) to the multi-label learning setting. CP complements the predictions of machine learning algorithms with reliable measures of confidence. As a result the proposed approach instead of just predicting the most likely subset of classes for a new unseen instance, also indicates the likelihood of each predicted subset being correct. This additional information is especially valuable in the multi-label setting where the overall uncertainty is extremely high.

1 Introduction

Most machine learning research on classification deals with problems in which each instance is associated with a single class y from a set of classes $\{Y_1, \ldots, Y_c\}$. As opposed to this standard setting, in *multi-label classification* each instance can belong to multiple classes, so that each instance is associated with a set of classes $\psi \subseteq \{Y_1, \ldots, Y_c\}$, called a *labelset*. There are many real-world problems in which such a setting is natural. For instance in the categorization of news articles an article discussing the positions of political parties on the educational system of a country can be classified as both politics and education. Although until recently the main multi-label classification application was the categorization of textual data, in the last few years an increasing number of new applications started to attract the attention of more researchers to this setting. Such applications include the semantic annotation of images and videos, the categorization of music into emotions, functional genomics, proteomics and directed marketing.

As a result of the increasing attention to the multi-label classification setting, many new machine learning techniques have been recently developed to deal with problems of this type, see e.g. [2, 8–11]. However, like most machine learning methods, these techniques do not produce any indication about the likelihood of each of their predicted labelsets being correct. Such an indication though can be very helpful in deciding how much to rely on each prediction, especially since

the certainty of predictions may vary to a big degree between instances. To address this problem this paper examines the application of a recently developed framework for providing reliable confidence measures to predictions, called *Conformal Prediction* (CP) [7], to the multi-label classification setting. Specifically it follows the newly proposed *Cross-Conformal Prediction* (CCP) [6] version of the framework, which allows it to overcome the prohibitively large computational overhead of the original CP. The proposed approach computes a p-value for each of the possible labelsets, which can be used either for accompanying each prediction with confidence measures that indicate its likelihood of being correct, or for producing sets of labelsets that are guaranteed to contain the true labelset with a frequency equal to or higher than a required level of confidence.

In the remaining paper, a description of the general idea behind CP and of the CCP version of the framework, is first provided in Section 2. Section 3 defines the proposed approach while Section 4 presents the experiments performed and the obtained results. Finally Section 5 gives the conclusions of this work.

2   Conformal and Cross-Conformal Prediction

Typically in classification we are given a set of training examples $\{z_1, \ldots, z_l\}$, where each $z_i \in \mathcal{Z}$ is a pair (x_i, y_i) consisting of a vector of attributes $x_i \in \mathbb{R}^d$ and the classification $y_i \in \{Y_1, \ldots, Y_c\}$. We are also given a new unclassified example x_{l+1} and our task is to state something about our confidence in each possible classification of x_{l+1} without assuming anything more than that all (x_i, y_i), $i = 1, 2, \ldots$, are independent and identically distributed.

The idea behind CP is to assume every possible classification Y_j of the example x_{l+1} and check how likely it is that the extended set of examples

$$\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_j)\} \tag{1}$$

is i.i.d. This in effect will correspond to the likelihood of Y_j being the true label of the example x_{l+1} since this is the only unknown value in (1).

First a function A called *nonconformity measure* is used to map each pair (x_i, y_i) in (1) to a numerical score

$$\alpha_i = A(\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_j)\}, (x_i, y_i)), \quad i = 1, \ldots, l, \tag{2a}$$

$$\alpha_{l+1}^{Y_j} = A(\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_j)\}, (x_{l+1}, Y_j)), \tag{2b}$$

called the *nonconformity score* of instance i. This score indicates how nonconforming, or strange, it is for i to belong in (1). In effect the nonconformity measure is based on a conventional machine learning algorithm, called the *underlying algorithm* of the corresponding CP and measures the degree of disagreement between the actual label y_i and the prediction $\hat{y}_i$ of the underlying algorithm, after being trained on (1). The nonconformity measure for multi-label learning used in this work is defined in Section 3.

The nonconformity score $\alpha_{l+1}^{Y_j}$ is then compared to the nonconformity scores of all other examples to find out how unusual (x_{l+1}, Y_j) is according to the

nonconformity measure used. This comparison is performed with the function

$$p((x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_j)) = \frac{|\{i = 1, \ldots, l, l+1 : \alpha_i^{Y_j} \geq \alpha_{l+1}^{Y_j}\}|}{l+1}, \quad (3)$$

the output of which is called the p-value of Y_j, also denoted as $p(Y_j)$. An important property of (3) is that $\forall \delta \in [0, 1]$ and for all probability distributions P on $\mathcal{Z}$,

$$P^{l+1}\{((x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1})) : p(y_{l+1}) \leq \delta\} \leq \delta; \quad (4)$$

a proof can be found in [7]. According to this property, if $p(Y_j)$ is under some very low threshold, say 0.05, this means that Y_j is highly unlikely as the probability of such an event is at most 5% if (1) is i.i.d.

There are two standard ways to use the p-values of all possible classifications for producing the output of a CP:

- Predict the classification with the highest p-value and output one minus the second highest p-value as confidence to this prediction and the p-value of the predicted classification (i.e. the highest p-value) as credibility.
- Given a confidence level $1 - \delta$, output the prediction set $\{Y_j : p(Y_j) > \delta\}$.

In the first case, confidence is an indication of how likely the prediction is of being correct compared to all other possible classifications, whereas credibility indicates how suitable the training set is for the particular instance; specifically a very low credibility value indicates that the particular instance does not seem to belong to any of the possible classifications. In the second case, the prediction set will not contain the true label of the instance with at most δ probability.

The important drawback of the above process is that since the last example in (1) changes for every possible classification, the underlying algorithm needs to be trained c times. Moreover the whole process needs to be repeated for every test example. This makes it extremely computationally inefficient in many cases and especially in a multi-label setting where there are 2^n possible labelsets for n classes, or $2^n - 1$ if we exclude the empty labelset.

To overcome this computational inefficiency problem an inductive version of the framework was proposed in [3] and [4] called *Inductive Conformal Prediction* (ICP). ICP is based on the same theoretical foundations described above, but follows a modified version of the approach, which allows it to train the underlying algorithm only once. This is achieved by dividing the training set into two smaller sets, the *proper training set* and the *calibration set*. The proper training set is then used for training the underlying algorithm of the ICP and only the examples in the calibration set are used for calculating the p-value of each possible classification for every test example.

Although ICP is much more computationally efficient than the original CP approach, the fact that it does not use the whole training set for training the underlying algorithm and for calculating its p-values results in lower informational efficiency. That is the resulting prediction sets might be larger than the ones produced by the original CP approach. Cross-Conformal Prediction, which

was recently proposed in [6], tries to overcome this problem by combining ICP with cross-validation. Specifically, CCP partitions the training set in K subsets (folds) $S_1, \ldots, S_K$ and calculates the nonconformity scores of the examples in each subset S_k and of (x_{l+1}, Y_j) for each possible classification Y_j as

$$\alpha_i = A(\cup_{m \neq k} S_m, (x_i, y_i)), \quad i \in S_k, \quad m = 1, \ldots, K, \tag{5a}$$

$$\alpha_{l+1}^{Y_j,k} = A(\cup_{m \neq k} S_m, (x_{l+1}, Y_j)), \quad m = 1, \ldots, K, \tag{5b}$$

where A is the given nonconformity measure. Note that for (x_{l+1}, Y_j) K nonconformity scores $\alpha_{l+1}^{Y_j,k}$, $k = 1, \ldots, K$ are calculated, one with each of the K folds. Now the p-value for each possible classification Y_j is computed as

$$p(Y_j) = \frac{\sum_{k=1}^{K} |\{(x_i, y_i) \in S_k : \alpha_i \geq \alpha_{l+1}^{Y_j,k}\}| + 1}{l + 1}. \tag{6}$$

The CCP version of the framework was chosen to be followed here due to its big advantage in computational efficiency over CP, since it needs to train the underlying algorithm only K times, and its advantage over ICP since it utilizes the whole training set for producing its p-values. As opposed to CP and ICP, the validity of which has been proven theoretically, at the moment there are no theoretical results about the validity of CCP. However in [6] its outputs have been shown to be empirically valid. The same is shown in the experimental results of this work, presented in Section 4.

3   ML-RBF Cross-Conformal Predictor

This section describes the proposed approach, which in effect comes down to the definition of a suitable nonconformity measure for multi-label classification and its use with the CCP framework. In order to use the CCP framework in the multi-label setting, the set of possible classifications $\{Y_1, \ldots, Y_c\}$ is replaced by the powerset $\mathcal{P}(\{Y_1, \ldots, Y_n\})$, where n is the number of the original classes of the problem. In the experiments that follow RBF Neural Networks for multi-label learning (ML-RBF) [8] is used as underlying algorithm as it seems to be one of the best performing algorithms designed specifically for multi-label problems. However the proposed approach is general and can be used with any other method which gives scores for each class.

After being trained on a given training set, for every test example ML-RBF produces a score for each possible class of the task at hand. It then outputs as its prediction the labelset containing all classes with score higher than zero; higher score indicates higher chance of the class to be in the labelset. In order to use the scores produced by ML-RBF for computing the nonconformity scores of the proposed Cross-Conformal Predictor, the former were transformed to the range $[0, 1]$ with the logistic sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}.$$

The nonconformity measure (5) for the multi-label CCP can now be defined based on the transformed outputs of ML-RBF for (x_i, ψ_i) after being trained on $\cup_{m \neq k} S_m$, $m = 1, \ldots, K$ as

$$\alpha_i = \sum_{j=1}^{n} |t_i^j - o_i^j|^d, \tag{7}$$

where n is the number of possible classes, t_i^j is 1 if $Y_j \in \psi_i$ and 0 otherwise, and o_i^j is the transformed output of the ML-RBF corresponding to class Y_j. Finally d is a parameter of the algorithm which controls the sensitivity of the nonconformity measure to small differences in o_i^j when $|t_i^j - o_i^j|$ is small as opposed to when it is large. This nonconformity measure takes into account the distance of all outputs from the true values. Note that in the case of test examples t_i^j is the corresponding value for each assumed labelset.

The nonconformity measure (7) takes into account only the outputs of the ML-RBF for each instance. However, it is very strange to have a pair of labels in a labelset that have never appeared together in the training set. So (7) was extended to take into account the occurrence of each pair of labels in the training set $\cup_{m \neq k} S_m$, $m = 1, \ldots, K$. This extended nonconformity measure is defined as:

$$\alpha_i = \sum_{j=1}^{n} |t_i^j - o_i^j|^d + \lambda \sum_{1 \leq j < r \leq n} t_i^j t_i^r \mu_{j,r}, \tag{8}$$

where $\mu_{j,r}$ is 0 if the labels Y_j and Y_r have been observed together in the labelset of at least one instance of the training set and 1 otherwise. In effect the additional part of this nonconformity measure adds λ to the nonconformity score of an example for each pair of labels in the labelset of the example ($t_i^j t_i^r = 1$) which have never been observed together in any instance of the training set ($\mu_{j,r} = 1$). The parameter λ adjusts the sensitivity to this part of the measure. A high value of λ makes this part of the measure dominate when a pair of labels that has never been observed before exists in the labelset of the example.

The complete proposed approach is derived by plugging in (8) as A in (5) and computing the p-value of each possible labelset with (6). The resulting p-values can be used in either of the two ways described in Section 2.

4 Experiments and Results

4.1 Data Sets

To evaluate the performance of the proposed approach two data sets from different application domains were used, one from the semantic scene analysis domain and one from the bioinformatics domain. The first data set, *scene* [1], is concerned with the semantic classification of pictures into one or more of the classes: beach, sunset, foliage, field, mountain and urban. It consists of 1211 training and 1196 test examples, each described by 294 features. The second data set, *yeast*

[2], is concerned with predicting the functional classes of genes in the Yeast Saccharomyces cerevisiae. Each gene is described by the concatenation of microarray expression data and a phylogenetic profile, and is associated with a set of 14 functional classes. The data set contains 1500 genes as training set and 917 genes as test set, each described by 103 features. Both data sets were obtained from the website of the Mulan library [5].

4.2   Single Prediction Evaluation

The first set of experiments evaluates the quality of the single predictions produced by the ML-RBF CCP and compares it to that of its underlying method and those of three other popular multi-label techniques, namely BP-MLL [10], ML-kNN [11] and ML-NB [9]. Four evaluation measures for multi-label classification were used. The first is Hamming Loss (HL), which is the most popular measure for multi-label problems, defined as:

$$HL = \frac{1}{g} \sum_{i=l+1}^{l+g} \frac{|\psi_i \triangle \hat{\psi}_i|}{n},$$ (9)

where $\{(x_{l+1}, \psi_{l+1}), \ldots, (x_{l+g}, \psi_{l+g})\}$ are the test examples, $\hat{\psi}_i$ is the predicted labelset for example i and $\triangle$ is the symmetric difference between two sets. The second measure is Classification Accuracy (CA) defined as:

$$CA = \frac{1}{g} \sum_{i=l+1}^{l+g} I(\psi_i = \hat{\psi}_i),$$ (10)

where $I(true) = 1$ and $I(false) = 0$. This measure is rather strict as it requires the predicted and true labelsets to be identical. The third and fourth measures are the *macro averaged* and *micro averaged* F-measure, which is the harmonic mean of precision and recall. The F-measure for a single label is defined as:

$$F(tp, tn, fp, fn) = \frac{2tp}{2tp + fp + fn},$$ (11)

where tp is the number of true positives, tn the number of true negatives, fp the number of false positives and fn the number of false negatives. In the multi-label case, if tp_j, tn_j, fp_j and fn_j are the same values for each label Y_j, then its macro averaged version is defined as:

$$F_{macro} = \frac{1}{n} \sum_{j=1}^{n} F(tp_j, tn_j, fp_j, fn_j).$$ (12)

The micro averaged version of the F-measure is defined as:

$$F_{micro} = F \left(\sum_{j=1}^{n} tp_j, \sum_{j=1}^{n} tn_j, \sum_{j=1}^{n} fp_j, \sum_{j=1}^{n} fn_j \right).$$ (13)

Table 1. Performance of the proposed approach and comparison to that of other multi-label algorithms on the scene data set.

Algorithm	Evaluation Measure			
	HL	CA	F_{macro}	F_{micro}
ML-RBF CCP with $\lambda = 0$	0.0928	0.6798	**0.7417**	**0.7363**
ML-RBF CCP with $\lambda = 1$	**0.0927**	**0.6831**	0.7410	0.7358
Orgiginal ML-RBF	0.0959	0.5468	0.6922	0.6890
BP-MLL	0.2903	0.1630	0.0509	0.1665
ML-kNN	0.0953	0.6012	0.7189	0.7183
ML-NB	0.1309	0.4105	0.6230	0.6221

Table 2. Performance of the proposed approach and comparison to that of other multi-label algorithms on the yeast data set.

Algorithm	Evaluation Measure			
	HL	CA	F_{macro}	F_{micro}
ML-RBF CCP with $\lambda = 0$	**0.1954**	0.1821	**0.3896**	**0.6432**
ML-RBF CCP with $\lambda = 1$	**0.1954**	0.1821	**0.3896**	**0.6432**
Orgiginal ML-RBF	0.1970	**0.1865**	0.3891	0.6407
BP-MLL	0.2272	0.0960	0.3047	0.6212
ML-kNN	0.1980	0.1658	0.3567	0.6360
ML-NB	0.2115	0.1254	0.3428	0.6152

Tables 1 and 2 present the performance of the proposed approach together with that of its underlying algorithm and that of BP-MLL, ML-kNN and ML-NB on the scene and yeast data sets respectively. The best values for each measure are highlighted in bold. For ML-RBF the fraction parameter was set to 0.01 and the scaling factor to 1 as in [8]. In the case of the CCP the number of folds for each data set was chosen so that each fold contained approximately 100 training instances, therefore 12 folds were used for the scene data set and 15 folds for the yeast dataset. The parameter d of the nonconformity measure (8) was set to 4, which seems to be a good choice based on the performed experiments, while for λ the two extreme values of 0 and 1 were used. Setting λ to 0 in effect corresponds to nonconformity measure (7) and setting it to 1 makes the nonconformity score of any labelset containing a pair of classes that has never been observed in the training set always higher than others. For BP-MLL the number of hidden neurons was set to 20% of the input dimensionality, the learning rate to 0.05 and the training epochs to 100 as in [10]. For ML-kNN the number of nearest neighbours was set to 10 and the smoothing parameter to 1 as in [11]. For ML-NB the percentage of remaining features after PCA was set to 0.3 as in [9].

The results in these tables show that not only the proposed approach provides important additional information about the likelihood of each of its predictions being correct, but it also outperforms its underlying algorithm and the three other popular multi-label techniques. The only exception is the classification

Table 3. Prediction set sizes and error rates at the 95%, 90% and 80% confidence levels for the scene data set.

# of labelsets	With $\lambda = 0$ Confidence Level			With $\lambda = 1$ Confidence Level		
	95%	90%	80%	95%	90%	80%
1	0.00%	2.34%	54.93%	6.77%	18.23%	62.63%
2	0.08%	25.50%	34.20%	9.62%	32.69%	29.43%
3 to 2^2	25.50%	65.64%	10.87%	43.31%	45.07%	7.94%
$(2^2 + 1)$ to 2^3	72.74%	6.52%	0.00%	40.05%	4.01%	0.00%
$(2^3 + 1)$ to 2^4	1.67%	0.00%	0.00%	0.25%	0.00%	0.00%
Errors	3.76%	9.28%	21.07%	3.60%	9.28%	20.99%

accuracy of the ML-RBF in the case of the yeast data set, but even in this case the accuracy of the proposed approach with both values of λ is very close. Comparing the performance of the ML-RBF CCP with the two different λ values one can see that in the case of the scene data there is only a negligible difference while in the case of the yeast data the performance remains the same for all evaluation measures. Therefore the value of λ does not have any important effect on the performance of the single predictions produced by the ML-RBF CCP. It does however affect the quality of the resulting p-values as will be shown in the next subsection.

4.3   Prediction Region Evaluation

The main advantage of the proposed approach over other multi-label techniques is the production of a p-value for each possible labelset of a new unseen instance, which can be translated either to confidence and credibility measures for its prediction or to prediction sets that are guaranteed to contain the true labelset at a required confidence level. This subsection examines the informativeness and reliability of the resulting prediction sets and consequently of the computed p-values and confidence measures. More specifically given a required level of confidence $1 - \delta$, the ML-RBF CCP produces a set of labelsets that has at most δ chance of not containing the true labelset of the unseen instance. The informativeness of this set of labelsets can be assessed in terms of its size, while its reliability can be assessed by the percentage of cases for which it does not contain the true labelset, this percentage should be less than or very near δ.

Tables 3 and 4 present the results of the proposed approach in this setting for the scene and yeast data sets respectively with the nonconformity measure parameter λ set to 0 and 1. The same parameters reported in Subsection 4.2 were used. The two tables report the sizes of the prediction sets produced for the 95%, 90% and 80% confidence levels together with the observed error percentages, i.e. the percentages of prediction sets that did not contain the true labelset.

Table 3 reports the results for the scene data set in terms of the percentage of prediction sets containing only 1, 2, 3 to 4, 5 to 8 and 9 to 16 labelsets for each

Table 4. Prediction set sizes and error rates at the 95%, 90% and 80% confidence levels for the yeast data set.

| # of | With $\lambda = 0$ | | | With $\lambda = 1$ | | |
| | Confidence Level | | | Confidence Level | | |
labelsets	95%	90%	80%	95%	90%	80%
$(2^6 + 1)$ to 2^7	0.00%	0.00%	1.42%	0.00%	0.00%	1.42%
$2^7 < l \leq 2^8$	0.00%	0.11%	3.71%	0.00%	0.22%	4.58%
$2^8 < l \leq 2^9$	0.55%	3.82%	14.07%	0.76%	4.36%	18.65%
$2^9 < l \leq 2^{10}$	3.05%	7.09%	60.96%	3.93%	10.14%	60.32%
$2^{10} < l \leq 2^{11}$	8.94%	34.46%	19.85%	13.85%	45.58%	15.05%
$2^{11} < l \leq 2^{12}$	38.71%	52.89%	0.00%	57.58%	39.48%	0.00%
$2^{12} < l \leq 2^{13}$	48.64%	1.64%	0.00%	23.88%	0.22%	0.00%
$2^{13} < l \leq 2^{14}$	0.11%	0.00%	0.00%	0.00%	0.00%	0.00%
Errors	4.69%	9.38%	19.85%	4.80%	9.60%	19.96%

confidence level; there was no prediction set containing more than 16 labelsets out of the possible 63. The last row of the table reports the percentage of errors observed for each confidence level. Comparing the results obtained with the nonconformity measure parameter λ set to 0 with those obtained with $\lambda = 1$, one can see that the latter produces much more informative prediction sets. Taking into account the classification accuracy of this data set (68.31%) and the large number of possible labelsets, the resulting prediction sets are quite tight. One can be 95% confident in about 60% of the test instances by considering less than 4 out of the possible 63 labelsets. By reducing the required confidence to 90% one can be certain in a single labelset for about 18% of the test instances and between one or two labelsets for about half the test instances. Finally with a confidence level of 80% a single labelset is given for more than 60% of the test instances. In terms of empirical reliability, only the percentages of errors for the 80% confidence level are slightly higher than the required significance level, which can be attributed to statistical fluctuations.

Table 4 presents the same results for the yeast data set. In this case the prediction sets contained a much higher number of labelsets so the table reports the percentage of prediction sets containing between $2^i + 1$ and 2^{i+1} labelsets with $i = 6, \ldots, 13$ for each confidence level; there were no prediction sets containing less than 2^6 labelsets. The rather big size of the resulting prediction sets is not strange baring in mind the very low classification accuracy of this data set, which is only 18.65%. Comparing the results obtained with the nonconformity measure parameter λ set to 0 with those obtained with $\lambda = 1$, again shows the superiority of nonconformity measure (8), as it produces smaller prediction sets. Considering the high difficulty of the particular task one can say that the resulting prediction sets are quite informative. The number of labelsets needed to satisfy the 80% confidence level is 1/16th or less ($\leq 2^{10}$) of all the possible labelsets for 85% of the test instances. Finally in terms of empirical reliability, the percentage of errors observed is in all cases below the required significance level.

5   Conclusions

This work examined the application of the conformal prediction framework to the multi-label setting. Unlike the other techniques developed for multi-label problems, the proposed approach accompanies each of its predictions with reliable measures of confidence. Experimental results on two popular multi-label data sets have shown that not only the proposed approach provides important additional information for each prediction, but it also outperforms other popular multi-label techniques. Furthermore its confidence measures have been shown to be informative and reliable. The provision of confidence measures can be very helpful in practical applications, considering the high uncertainty that exists in this setting.

Future work includes the development of additional nonconformity measures and the experimentation with more multi-label data. In addition generating separate p-values for each class and combining them for obtaining the p-value of each labelset could also be examined as an alternative. Finally the possibility of generating a ranking of the possible classes for each instance would also be a good addition to multi-label CP.

References

1. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. Pattern Recognition 37, 1757–1771 (2004)
2. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Information Processing Systems 14. pp. 681–687. MIT Press (2002)
3. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Proceedings of the 13th European Conference on Machine Learning (ECML'02). LNCS, vol. 2430, pp. 345–356. Springer (2002)
4. Papadopoulos, H., Vovk, V., Gammerman, A.: Qualified predictions for large data sets in the case of pattern recognition. In: Proceedings of the 2002 International Conference on Machine Learning and Applications (ICMLA'02). pp. 159–163. CSREA Press (2002)
5. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. pp. 667–685 (2010)
6. Vovk, V.: Cross-conformal predictors. Annals of Mathematics and Artificial Intelligence (2013), http://dx.doi.org/10.1007/s10472-013-9368-4
7. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
8. Zhang, M.L.: ML-RBF: RBF neural networks for multi-label learning. Neural Processing Letters 29(2), 61–74 (2009)
9. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences 179(19), 3218–3229 (2009)
10. 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)
11. Zhang, M.L., Zhou, Z.H.: ML-kNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)