Proceedings of Machine Learning Research 230:1–14, 2024 Conformal and Probabilistic Prediction with Applications

Multi-label Conformal Prediction with a Mahalanobis Distance Nonconformity Measure

Kostas Katsios K.KATSIOS@ALBOURNE.COM
Harris Papadopoulos H.PAPADOPOULOS@FREDERICK.AC.CY
Computational Intelligence Research Lab.,
Frederick University, Nicosia, Cyprus

Machine Learning Research Group,
Albourne Partners Ltd, London, UK

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

This preliminary study introduces a Conformal Prediction method for Multi-label Classification with a nonconformity measure based on the Mahalanobis distance. The Mahalanobis measure incorporates a covariance matrix considering correlations between the errors of the underlying classifier on each label. Our experimental results show that this approach results in a significant informational efficiency improvement over the previously proposed Euclidean Norm nonconformity measure.

Keywords: Mahalanobis, Multi-label, Classification, Conformal, Prediction, Confidence, Power-set, Euclidean

1. Introduction

Multi-label classification is a problem category in which each instance can belong to multiple classes simultaneously, resulting in the formation of label-sets. The complexity of such tasks arises from the need to consider numerous or all possible combinations of classes. Multi-label learning algorithms use various techniques to predict these label-sets. The primary categories of these techniques include Problem Transformation (PT) and Algorithm Adaptation (AA) methods. Transformation methods, such as binary relevance or classifier chains, decompose the multi-label classification problem into binary classification tasks for each class. Conversely, algorithm adaptation methods correspond to modified versions of multi-class machine learning techniques for predicting sets of labels, see (Tsoumakas and Katakis, 2007) and (Zhang and Zhou, 2013) for comprehensive reviews.

The focus of this work is on the reliable quantification of uncertainty of Multi-label learning through Conformal Prediction (CP). Specially, we propose an Inductive (or Split) Conformal Prediction (ICP) approach that can be combined with any classifier that produces a score for each class. The proposed approach provides prediction regions of label-sets with guaranteed $1-\varepsilon$ coverage of the true label-set, for any required significance level ε.

A number of CP techniques have been proposed so far for multi-label classification. One such technique, referred to as Label Power-set (LP) CP, was proposed by Papadopoulos (2014). As its name suggests, this technique assigns a p-value to each possible label-set. This is done through a nonconformity measure that calculates the sum of absolute differences between the probabilistic outputs of the underlying model and the actual labels. An

extension to this nonconformity measure is also defined that considers the co-occurrences of labels in label-sets, leading to tighter prediction regions. The LP approach was recently extended in (Maltoudoglou et al., 2022) with the proposal of an approach for the efficient computation of the prediction regions based on a Euclidean Norm nonconformity measure.

Another technique is Instance Reproduction (IR), as presented in the work by Wang et al. (2014). The authors utilize a problem transformation algorithm to create a new single-label dataset. They then apply single-label machine learning algorithms to measure the confidence level of each label and select multiple labels whose p-values are greater than the significance level. Additionally, Wang et al. (2015), use the Binary Relevance (BR) technique to provide prediction regions based on separate binary classifiers for each class. This method guarantees the composition of valid prediction regions using the Bonferroni inequality. Similarly, Lambrou and Papadopoulos (2016) propose a generalization of the Binary Relevance CP based on Hamming loss metric.

A recent work in this field by Tyagi and Guo (2023) introduces a method for constructing a hierarchical tree for multi-label classification using the technique of multiple hypothesis testing. In this hierarchical testing procedure, confidence evaluation occurs by applying Bonferroni correction at each tree layer.

Our study extends the work of Maltoudoglou et al. (2022) and is inspired by the work of Messoudi et al. (2022) who proposed a Mahalanobis distance nonconformity measure for Multi-target Regression tasks. Mahalanobis distance represents a transformation of Euclidean distance through a covariance matrix, derived from the proper-training data, considering correlations between error vectors. By employing the Mahalanobis nonconformity measure, we observe that the predicted sets are smaller compared to those obtained using the Euclidean norm. This improvement is evident in our experimental results.

The rest of this paper starts with an overview of ICP in Section 2. This is followed by the definition of the proposed approach in Section 3. Section 4 details the experimental evaluation of our approach and its comparison with the previously proposed Euclidean Norm approach. Finally, Section 5 gives our conclusions and plans for future work.

2. Inductive Conformal Prediction (ICP)

Inductive Conformal Prediction (ICP) was introduced in (Papadopoulos et al., 2002a) and (Papadopoulos et al., 2002b) to address the computational inefficiency issue of Full CP. Let X denote the feature space and Ψ denote a set of labels. The inputs to the feature space are represented as vectors of the form, $\boldsymbol{x}_i = (x_{i_1}, ..., x_{i_k}, ..., x_{i_s})$, for $k = 1, ..., s$, where $X \cong \mathbb{R}^s$ and s is the number of attributes. Pairing the elements of the feature space X and the set Ψ forms the example space, $Z = \{(x_i, \psi_i) \ : \ x_i \in X, \psi_i \in \Psi\}$, for $i = 1, ..., n$.

The training set is split in two parts, the proper-training set $\{(x_1, \psi_1), ..., (x_q, \psi_q)\}$, where $q \leq n$, and the calibration set $\{(x_{q+1}, \psi_{q+1}), ..., (x_n, \psi_n)\}$. Additionally, we define a nonconformity measure $A : Z \to \mathbb{R}$ which assigns a score that describing how different an instance x_{n+m} is from the instances in the proper-training set. The underlying model is trained on the proper-training set, and the training results are used to calculate the nonconformity scores of the calibration instances as,

$$a_i = A\Big(\{(x_1, \psi_1), ..., (x_q, \psi_q)\}, (x_i, \psi_i)\Big), \ i = q + 1, ..., n. \tag{1}$$

The objective is to produce a set-prediction $\Gamma^{\varepsilon}_{n+m} \subseteq \Psi$ that contains the true targets with probability $1-\varepsilon$, where ε is the significance level, for every test instance x_{n+m}. The nonconformity score for every assumed label $\mathcal{Y}_j$ is given by,

$$a^{\mathcal{Y}_j}_{n+m} = A\Big(\big\{(x_1,\psi_1),...,(x_q,\psi_q)\big\},(x_{n+m},\mathcal{Y}_j)\Big). \tag{2}$$

The nonconformity scores of calibration instances are used to calculate the p-value p of each possible label $\mathcal{Y}_j$,

$$p(\mathcal{Y}_j) = \frac{\big|i=q+1,...,n \ : \ a_i \geq a^{\mathcal{Y}_j}_{n+m}\big|+1}{n-q+1} \tag{3}$$

The conformal prediction regions for every test instance x_{n+m} are defined as,

$$\Gamma^{\varepsilon}_{x_{n+m}} = \big\{\mathcal{Y}_j : p(\mathcal{Y}_j) > \varepsilon\big\}. \tag{4}$$

Instead of calculating the p-value for every possible label $\mathcal{Y}_j$, we can find a threshold on the test instance nonconformity score, below which the p-value will be greater than a given significance level. We sort the calibration scores in descending order and we denote the ordered calibration scores as a^{desc}_k, for $k=1,...,n-q$, where $a^{desc}_1 < ... < a^{desc}_{n-q}$. Therefore, for any given significance level ε, we find the minimum integer $k_\varepsilon \in \{1,...,n-q\}$ that satisfies,

$$\Big|\big\{i=q+1,...,n \ : \ a^{desc}_i \geq a^{desc}_{k_\varepsilon}\big\}\Big| > \varepsilon(n-q+1)-1. \tag{5}$$

Since the calibration scores are sorted in descending order, the number of scores satisfying the inequality (5) is,

$$k_\varepsilon = \Big|\big\{i=1,...,n-q \ : \ a^{desc}_i \geq a^{desc}_{k_\varepsilon}\big\}\Big|, \tag{6}$$

and the inequality can be written as,

$$k_\varepsilon > \varepsilon(n-q+1)-1. \tag{7}$$

Lemma 1 *For any significance level $\varepsilon \in [0,1]$ and the numbers $n,q \in \mathbb{Z}$, we have,*

$$\lfloor \varepsilon(n-q+1) \rfloor = \begin{cases} \lceil \varepsilon(n-q+1) \rceil, & \text{if } \varepsilon(n-q+1) \in \mathbb{Z} \\ \lceil \varepsilon(n-q+1)-1 \rceil, & \text{if } \varepsilon(n-q+1) \notin \mathbb{Z} \end{cases} \tag{8}$$

Proof *Based on the definitions of the floor and ceil functions, it implies that,*

$$\lceil \varepsilon(n-q+1) \rceil - \lfloor \varepsilon(n-q+1) \rfloor = \begin{cases} 0, & \text{if } \varepsilon(n-q+1) \in \mathbb{Z} \\ 1, & \text{if } \varepsilon(n-q+1) \notin \mathbb{Z}. \end{cases}$$

The first case of equality (8) is clear. The second case is proven by using the property,

$$\lceil \varepsilon(n-q+1) \rceil - 1 = \lceil \varepsilon(n-q+1)-1 \rceil.$$

■

Proposition 2 *For some value ε of the significance level, the minimum integer of which the inequality (5) holds is,*

$$k_\varepsilon = \lfloor \varepsilon(n-q+1) \rfloor. \tag{9}$$

Proof *The integer k_ε is the minimum that satisfies inequality (9). From the definition of ceil function, it follows that,*

$$\lceil \varepsilon(n-q+1)-1 \rceil = \min\{k_\varepsilon \in \mathbb{Z} \ : \ k_\varepsilon > \varepsilon(n-q+1)-1\}.$$

The equality (9) is proven by applying Lemma (1). ■

Given k_ε, the prediction sets for each instance x_{n+m} at the ε significance level are written in the equivalent form,

$$\Gamma^\varepsilon_{x_{n+m}} = \{\mathcal{Y}_j \ : \ a^{\mathcal{Y}_j}_{n+m} \leq a^{desc}_{k_\varepsilon}\}. \tag{10}$$

3. Multi-label ICP using Mahalanobis measure

This section outlines the inductive conformal prediction approach, focusing on the introduction of the Mahalanobis nonconformity measure for multi-label classification defined in a vector space. First, we describe the multi-hot vector representation used in our notation. Next, we define the vector error space in which the ICP is to be performed. Finally, we define the Mahalanobis nonconformity measure and our algorithm.

3.1. Multi-hot label representation

Let $C = \{c_1, ..., c_d\}$ denote the set of d individual classes, with each class indexed corresponding to an element of C. A label-set $\mathcal{Y}_j$ is a subset of C. Hence, we define the power-set as,

$$\mathcal{P}(C) = \{\mathcal{Y}_j \ : \ \mathcal{Y}_j \subseteq C\}$$

which contains all possible label-sets $\mathcal{Y}_j$, $j = 1, ..., w$, generated by all combinations of class indices C, where w represents the total number of possible label-sets for d different classes and is calculated as $w = 2^d$.

In the subsequent sections we define Mahalanobis nonconformity measure in terms of the vector of errors between predicted probabilities and target label-sets in Euclidean vector space. To maintain consistency with vector space terminology, we convert the label-sets ψ_j into multi-hot vectors. For every label-set $\mathcal{Y}_j \in \mathcal{P}(C)$, we construct a multi-hot vector $\boldsymbol{y}_j = (y_{j_1}, ..., y_{j_c}, ..., y_{j_d})$ as follows,

$$y_{j_c} = \begin{cases} 0, \text{ if } c \notin \mathcal{Y}_j \\ 1, \text{ if } c \in \mathcal{Y}_j \end{cases}, \text{ for every } c \in C.$$

Note that the empty set in $\mathcal{P}(C)$ corresponds to the zero vector. Consequently, we create a bijection, $\sigma : \mathcal{P}(C) \to Y$, between the power-set $\mathcal{P}(C)$ and the formed subspace $Y \subseteq \mathbb{R}^d$ of the vectors $\boldsymbol{y}_j$. The number of possible multi-hot vectors in Y equals the number w of possible label-sets in $\mathcal{P}(C)$. Moreover, the true label of an instance i is included in the power-set $\mathcal{P}(C)$ and consequently in space Y. To distinguish it from the possible multi-hot vectors $\boldsymbol{y}_j$, we represent it as the multi-hot vector $\boldsymbol{t}_i = (t_{i_1}, ..., t_{i_d})$.

3.2. Error space

In the Euclidean vector space $\mathbb{R}^d$ we represent the predicted probabilities of classifier, for an instance x, and the multi-hot representation of the label-sets with vector structure, denoted as $\boldsymbol{o} = \boldsymbol{o}(x)$ and $\boldsymbol{y}$ respectively. We define the linear transformation $r : \mathbb{R}^d \times \{\boldsymbol{o}(x)\} \to \mathbb{R}^d$ with,

$$r(\boldsymbol{y}, \boldsymbol{o}(x)) = |\boldsymbol{y} - \boldsymbol{o}(x)|, \tag{11}$$

where we calculate the difference between predicted probabilities of the classifier and label-sets in vector form. We then give the definition of the error vectors $\boldsymbol{r}_i^{y_j}$ referring to a label-set $\boldsymbol{y_j}$, for an instance i.

Definition 3 *We define* $\boldsymbol{r}_i^{y_j} = \left(r_{i_1}, ..., r_{i_d}\right)$ *as the error vector for instance* i *related to label-set* y_j*, such that*

$$\boldsymbol{r}_i^{y_j} = (|y_{j_1} - o_{i_1}|, ..., |y_{j_d} - o_{i_d}|), \tag{12}$$

where $\boldsymbol{o_i} = (o_{i_1}, ..., o_{i_d})$*, with* $o_{i_k} \in [0, 1]$ *and* $k = 1, ..., d$*.*

The error vectors constitute a subspace R of $\mathbb{R}^d$. Given the predicted probabilities of an instance x, the linear map is defined as $r : Y \times \{\boldsymbol{o}(x)\} \to R$. The map is injective, and thus the label-space Y and the error space R are isomorphic.

3.3. Mahalanobis nonconformity measure

In multi-label classification, each instance belongs to multiple classes. Thus, the true classification ψ_{n+m} of an instance x_{n+m} is contained in the power-set $\mathcal{P}(C)$. According to vector space representation, the true label-set for the instance x_{n+m} is denoted as t_{n+m}. The conformal prediction approach provides the guarantee for an instance x_{n+m} that

$$\mathbb{P}\Big(t_{n+m} \in \Gamma^{\varepsilon}_{x_{n+m}} \ : \ p(t_{n+m}) > \varepsilon\Big) \geq 1 - \varepsilon \tag{13}$$

This guarantee is provided by the approaches proposed in (Lambrou and Papadopoulos, 2016) and (Papadopoulos, 2014). Based on the one-to-one correspondence between $\mathcal{P}(C)$ and subspace Y, it therefore holds that $\Gamma^{\varepsilon}_{x_{n+m}} \subseteq \mathcal{P}(C)$.

The choice of defining error vectors in Euclidean vector space provides a connection between the probabilistic outputs of the underlying classifier and the label-sets. Additionally, the Euclidean distance (or norm) establishes a relationship between errors with similar behavior. Thus, Maltoudoglou et al. (2022) define a nonconformity measure for multi-label classification using the Euclidean Norm, which for an instance i is expressed as,

$$\alpha_i^{y_j} = \sqrt{r_{i_1}^2 + ... + r_{i_d}^2}, \tag{14}$$

where $\boldsymbol{y_j}$ is the true label $\boldsymbol{t_i}$ for calibration instances and the assumed label for test instances. As mentioned, the Mahalanobis distance is a transformation of the Euclidean distance achieved by using the covariance matrix, denoted as Σ, which is symmetric and positive definite.

In the following, we define the Mahalanobis nonconformity measure for multi-label classification.

Definition 4 *Based on the Mahalanobis distance, we define the non-conformity measure of the error vectors for a calibration instance i as,*

$$\alpha_i^{t_i} = \sqrt{\left(\boldsymbol{r}_i^{t_i}\right)^T \ \Sigma^{-1} \ \boldsymbol{r}_i^{t_i}}, \tag{15}$$

where Σ^{-1} is the inverse covariance matrix which is estimated from error vectors of the proper training data.

Accordingly, we define the non-conformity measure of the error vectors for a possible label-set $\boldsymbol{y_j}$ of a test instance i as,

$$\alpha_i^{y_j} = \sqrt{\left(\boldsymbol{r}_i^{y_j}\right)^T \ \Sigma^{-1} \ \boldsymbol{r}_i^{y_j}}. \tag{16}$$

Lemma 5 *Given a fixed covariance matrix Σ and let $\boldsymbol{r}_{q+1}^{t_{q+1}}, ..., \boldsymbol{r}_n^{t_n}, \boldsymbol{r}_{n+m}^{t_{n+m}}$ be exchangeable error vectors. Then, the nonconformity scores $a_{q+1}^{t_{q+1}}, ..., a_n^{t_n}, a_{n+m}^{t_{n+m}}$ are also exchangeable.*

Proof *Since $\boldsymbol{r}_{q+1}^{t_{q+1}}, ..., \boldsymbol{r}_n^{t_n}, \boldsymbol{r}_{n+m}^{t_{n+m}}$ are exchangeable, for any permutation function $\pi : [n] \to [n]$, then for Mahalanobis nonconformity measure we have,*

$$a_{\pi(i)}^{t_{\pi(i)}} = \sqrt{\left(\boldsymbol{r}_{\pi(i)}^{t_{\pi(i)}}\right)^T \ \Sigma^{-1} \ \boldsymbol{r}_{\pi(i)}^{t_{\pi(i)}}} = \sqrt{\left(\boldsymbol{r}_i^{t_i}\right)^T \ \Sigma^{-1} \ \boldsymbol{r}_i^{t_i}},$$

for every $i = q+1, ..., n, n+m$. Thus, the measure (15) preserves exchangeability. ■

By the following theorem we prove the validity of the conformal predictor associated with the Mahalanobis nonconformity measure.

Theorem 6 *Given the exchangeable error vectors $\boldsymbol{r}_1^{t_1}, ..., \boldsymbol{r}_n^{t_n}$ and significance level $\varepsilon \in [0, 1]$, then for the Mahalanobis nonconformity measure it holds that,*

$$\mathbb{P}\Big(t_{n+m} \in \Gamma_{x_{n+m}}^{\varepsilon} \ : \ p(t_{n+m}) > \varepsilon\Big) \geq 1 - \varepsilon,$$

where x_{n+m} is a new instance.

Proof *Let $t_{n+m} \in \Gamma_{x_{n+m}}^{\varepsilon}$. This is true if and only if $a_{n+m}^{t_{n+m}} \leq a_{k_\varepsilon}^{desc}$, as defined in (10). Since error vectors are exchangeable, then Mahalanobis nonconformity scores inherit the exchangeability property (see Lemma 5). The probability function is given by,*

$$\mathbb{P}\Big(t_{n+m} \in \Gamma_{x_{n+m}}^{\varepsilon} \ : \ a_{n+m}^{t_{n+m}} \leq a_{k_\varepsilon}^{desc}\Big) \geq 1 - \varepsilon.$$

Since all permutations are equiprobable, then

$$\mathbb{P}\Big(t_{n+m} \in \Gamma_{x_{n+m}}^{\varepsilon} \ : \ p(t_{n+m}) > \varepsilon\Big) \geq 1 - \varepsilon.$$

■

The next algorithm outlines the steps involved in using the Mahalanobis nonconformity measure for Multi-label ICP.

Algorithm 1: Multi-label ICP using Mahalanobis measure

Input:

- Classifier' s predicted probabilities for proper-training data $\boldsymbol{o}(x_i)$, $i = 1, ..., q$, for calibration data $\boldsymbol{o}(x_i)$, $i = q + 1, ..., n$, for each test instance $o(x_{n+m})$.
- Label-sets of proper-training data $\boldsymbol{t}_i$, $i = 1, ..., q$, of calibration data $\boldsymbol{t}_i$, $i = q + 1, ..., n$.
- Required significance level ε.

1. Preprocessing on proper-training data:
   - Calculate the error vectors $\boldsymbol{r}_i = |\boldsymbol{o_i} - \boldsymbol{t_i}|$, $i = 1, ..., q$.
   - Form the covariance matrix Σ.
2. Preprocessing on calibration data:
   - Calculate the calibration nonconformity scores a_i, $i = q + 1, ..., n$, using (15).
   - Sort calibration scores in descending order a_k^{desc}, $k = 1, ..., n - q$.
   - Calculate k_ε using (9).
3. Calculate scores $a_{n+m}^{y_j}$, for every possible label-set $\boldsymbol{y_j} \in Y$, using (16).

Output: Predicted set, $\Gamma_{x_{n+m}}^{\varepsilon} = \left\{\boldsymbol{y}_j \in Y \ : \ a_{n+m}^{y_j} \leq a_{k_\varepsilon}^{desc}\right\}$.

4. Experiments

4.1. Datasets and Underlying Classifier

To evaluate the efficiency of the Mahalanobis nonconformity measure relative to the Euclidean Norm nonconformity measure, we employ two datasets with distinct properties. The size of a power-set affects the computational cost and time for producing all label-sets combinations. The Emotions and Yeast dataset are widely recognized for multi-label classification tasks. Table 1 provides detailed information on the datasets, including the number of instances, attributes, labels, and cardinality.

Table 1: Datasets for multi-label classification

Dataset	**Instances**	**Attributes**	**Labels**	**Cardinality**
Emotions	593	72	6	1.868
Yeast	2417	103	14	4.237

The underlying classifier is a Multi-layer Perceptron (MLP) model, with multiple five fully connected layers, a single dropout layer and batch normalization layer. Activation function relu is defined in each layer and the sigmoid activation function is set up for the probabilistic outputs.

Our experiments were performed following a 10-fold cross-validation process, which was repeated 10 times. The results were calculated as the average over all folds and repetitions.

The training set of each fold was further partitioned into a proper training set, validation set, and calibration set. The validation set was used for early stopping. Table 2 shows the number of instances allocated to each of these sets for each dataset.

Table 2: Dataset partitioning

	Proper train	Validation	Calibration	Test
Emotions	354	81	99	59
Yeast	1293	327	555	242

4.2. Forced prediction

In order to allow the comparison of the performance of the proposed approach with that of the underlying model using the typical multi-label classification metrics, we employ the forced prediction mode of CP. In this mode the CP outputs a single label-set prediction, corresponding to the highest p-value. This prediction is associated with a confidence score, indicating how likely is the predicted classification compared to all other possible classifications and a credibility value, measuring how common the test instance is compared to the proper-training set. A low credibility value indicates the instance is strange for all classes and differs from calibration instances. Let $\boldsymbol{z_i} = (z_{i_1}, ..., z_{i_d})$ be the multi-hot representation of the forced predicted label-set, the typical multi-label classification metrics we use defined as follows:

- Classification accuracy (CA) is calculated for the whole test set. It is a strict metric, since a correct prediction is given if the forced prediction matches with the true label represented as multi-hot vector $\boldsymbol{t_i} = (t_{i_1}, ..., t_{i_d})$, for every test instance $i = 1, ..., g$. The CA is defined as,

$$CA = \frac{1}{g} \sum_{i=1}^{g} I(\boldsymbol{t_i} = \boldsymbol{z_i}), \tag{17}$$

  where $I(true) = 1$ and 0 otherwise.

- The F1-measure corresponds to the harmonic mean of precision and recall and its value is in the range [0, 1]. In the multi-label case, the combined F1-measure over the labels can be calculated in two ways:

  - F1-micro is averaged over the complete set of test instances, which means that frequent labels weight more than infrequent ones. It is given by,

$$F1 - micro = \frac{2 \sum_{j=1}^{d} \sum_{i=1}^{g} t_{i_j} z_{i_j}}{\sum_{j=1}^{d} \sum_{i=1}^{g} t_{i_j} + \sum_{j=1}^{d} \sum_{i=1}^{g} z_{i_j}} \tag{18}$$

– F1-macro is averaged first per instance and the results are then averaged over the total number of labels. Consequently, F_{macro} gives equal weights to all labels and it therefore tends to be lower than F_{micro} when poorer performance is observed for the more infrequent ones. It is written as,

$$F1 - macro = \frac{1}{d}\sum_{j=1}^{d}\frac{\sum_{i=1}^{g} t_{i_j} z_{i_j}}{\sum_{i=1}^{g} t_{i_j} + \sum_{i=1}^{g} z_{i_j}} \tag{19}$$

- Hamming Loss (HL) averages the number of wrong labels over the total number of labels. It is defined as,

$$HL = \frac{1}{gd}\sum_{j=1}^{d}\sum_{i=1}^{g} t_{i_j} \oplus z_{i_j}, \tag{20}$$

where $\oplus$ indicates xor operator.

The Average-Confidence ($\overline{Conf}$) and Average-Credibility $\overline{Cred}$ are indicative of the performance in the set-prediction mode of Conformal Prediction:

- Average-Confidence ($\overline{Conf}$) is intended as an overall indication of how likely predictions are compared to all other possible classifications. It is written as,

$$\overline{Conf} = \frac{1}{g}\sum_{i=1}^{g} 1 - \max_{\boldsymbol{y_j} \neq \arg\max p_i(\boldsymbol{y_j})} p_i(\boldsymbol{y_j}) \tag{21}$$

where we compute the average value of all confidence scores (i.e. 1−the second largest p-value, over all considered label-sets $\boldsymbol{y_j}$) over g number of test instances.

- Average-Credibility ($\overline{Cred}$) is an overall metric which indicates how suitable is the training dataset for each test instance. It is defined as,

$$\overline{Cred} = \frac{1}{g}\sum_{1}^{g} \max_{\boldsymbol{y_j}} p_i(\boldsymbol{y_j}) \tag{22}$$

where the credibility of example i is the largest p-value out of all considered label-sets $\boldsymbol{y_j}$.

Tables 3 and 4, for the emotions and yeast dataset respectively, present the scoring results of the underlying classifier for the Accuracy, F1-micro, F1-macro, Hamming loss metrics and compare them with the performance metrics of the forced predictions Average-Confidence and Average-Credibility using the Mahalanobis and Euclidean norm nonconformity measure. F1-micro is the average of true positives and false negative and positives predicted label-sets. Also, does not consider the proportion of each class in the dataset. So, it reflect the accuracy on imbalanced data. In opposite, the F1-macro does to take label imbalance into account.

For the emotions dataset, we observe that the Hamming loss is 0.329, Accuracy is 0.04, F1-micro score is 0.226 and F1-macro is 0.103. Despite having a relatively small number of

classes (six), the underlying classifier performed poorly, with a low F1-micro score indicating worse performance on frequent label sets than on infrequent ones. Also, the size of predicted regions is affected by the classifier's performance, as indicated by the low accuracy score. The results of ICP are very close to the ones of the original MLP-classifier, while also providing the additional confidence and credibility information.

Table 3: Emotions dataset - Performance metrics

	MLP-classifier	ICP-Mahalanobis	ICP-Norm
Hamming loss	0.329	0.343	0.343
Accuracy	0.040	0.039	0.039
F1 Micro	0.226	0.246	0.246
F1 Macro	0.103	0.123	0.123
Average confidence	-	0.080	0.067
Average credibility	-	0.948	0.958

For the yeast dataset we report Hamming loss 0.198, Accuracy 0.186, F1-micro score 0.644 and F1-macro 0.38. The classifier performs better on this dataset, which has 14 classes and a significantly larger power-set. Moreover, the performance of forced-prediction was negligibly different from that of the underlying classifier.

Table 4: Yeast dataset - Performance metrics

	MLP-classifier	ICP-Mahalanobis	ICP-Norm
Hamming loss	0.198	0.200	0.200
Accuracy	0.186	0.158	0.158
F1 Micro	0.644	0.628	0.628
F1 Macro	0.380	0.336	0.336
Average confidence	-	0.203	0.205
Average credibility	-	0.851	0.822

The underlying classifier is trained on the whole training set and the ICP algorithm on the proper-training set. The performance results indicate that no substantial classification performance is sacrificed by the use of ICP. We conclude that, for the two datasets, the confidence information given by $\overline{Conf}$ is high. The Average-Confidence metric provides the probabilities that the predicted label-set is the true target. In addition, the high $\overline{Cred}$ points out that the proper-training data are suitable for classifying the test instances.

4.3. Statistical efficiency

In this Section we evaluate the statistical efficiency of the p-values and prediction regions produced by CP. For this, we use two of the probabilistic efficiency criteria proposed in (Vovk et al., 2016):

- $S-criterion$ is applicable on all possible p-values, $(p_i^{\boldsymbol{y_j}} \ : \ \boldsymbol{y_j} \in Y)$, for every test instance,

$$\frac{1}{g}\sum_{i=1}^{g}\sum_{\boldsymbol{y_j}} p_i^{\boldsymbol{y_j}}. \tag{23}$$

  It is independent of significance level ε. Small values are preferable in efficiency comparison.

- $N-criterion$ calculates the average number of predicted label-sets of all test instances, for significance level ε. It is defined as,

$$\frac{1}{g}\sum_{i=1}^{g}|\Gamma_i^{\varepsilon}|, \tag{24}$$

  where $|\Gamma_i^{\varepsilon}|$ is the size of the predicted region for each instance i at a significant level ε. Small values are preferable.

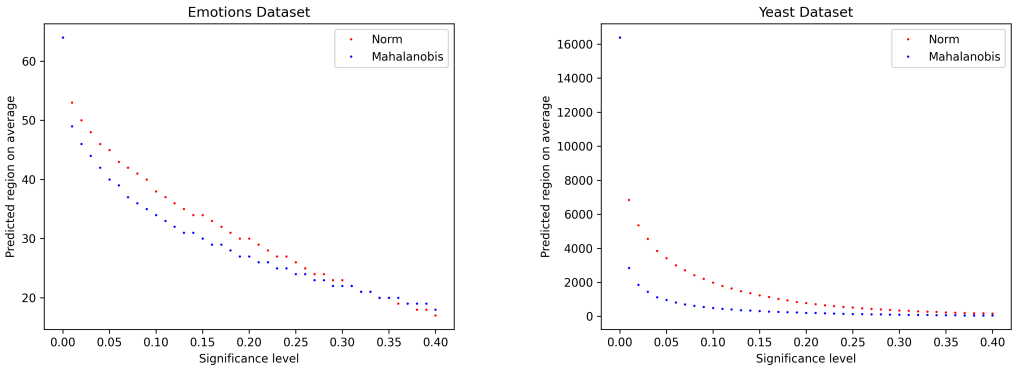
First, we report the values of the S criterion to measure efficiency by the average sum of the p-values obtained from the Mahalanobis and Norm nonconformity measures for both the Emotions and Yeast datasets. The results are presented in the Table 5. For the Emotions dataset, the $S-criterion$ value provided by the p-values using the Mahalanobis measure is slightly smaller than the value associated with the Norm measure. However, for the Yeast dataset the $S-criterion$ values are significantly different. The Mahalanobis nonconformity measure generates significantly smaller p-values overall.

Table 5: Mahalanobis and Norm S-criterion comparison

	Mahalanobis	Norm
Emotions	547.005	560.869
Yeast	30922.511	81839.323

Figure 1 presents the comparison of prediction region size on average per significance level ε between the Mahalanobis and Norm nonconformity measures. Particularly, we focus on significance levels in the range $[0, 40]$. For both datasets, the conformal predictor associated with the Mahalanobis nonconformity measure gives on average smaller prediction regions than the ones associated with the Norm measure.

We note that the number of possible label-sets is 64 and 16.384 for the Emotions and Yeast dataset, respectively. In Table 6, we focus on four significance level values: 0.01, 0.05, 0.1, 0.2 and report the prediction region size as a percentage of the number of possible label-sets. In all cases the Mahalanobis measure produces smaller regions, with the values for the Yeast dataset demonstrating an impressive reduction. In the best case, of the 0.2 significance level, the mean prediction region size of the Mahalanobis measure is five times smaller.



(a) (b)

Figure 1: Mahalanobis and Norm N-Criterion - Graph comparison.

Table 6: Mean Prediction Region size as a percentage of the number of possible label-sets

Emotions dataset			Yeast dataset		
Level	Mahala (%)	Norm (%)	Level	Mahala (%)	Norm (%)
0.01	77	83	0.01	17	42
0.05	62	70	0.05	6	21
0.10	53	59	0.10	3	12
0.20	42	47	0.20	1	5

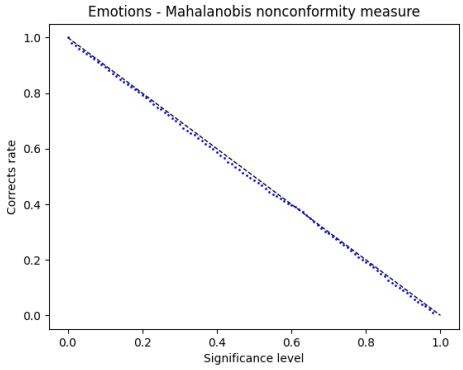
4.4. Empirical coverage

Figures 2 and 3 display the correct-rate of the predicted sets (percentage that covers the true label-set) obtained per significance level in the range of $[0, 1]$ for the two datasets. For the both datasets, the correct-rate closely aligns with the diagonal line, indicating a strong correspondence between the nominal and empirical coverage rates.

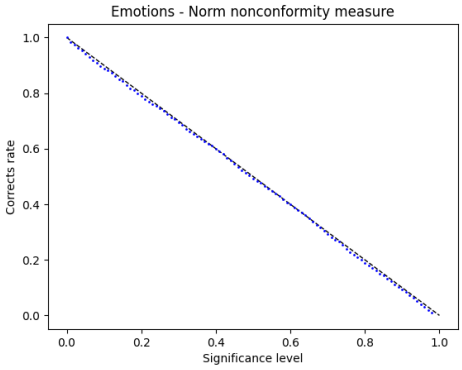
5. Conclusions and future work

This work proposes a multi-label ICP with Mahalanobis distance nonconformity measure. In particular, the Mahalanobis transformation is defined by a covariance matrix formed by error vectors in the proper-training set. These vectors in the error space are injectively mapped to the label-sets space, rendering the conformal predictor associated with the Mahalanobis measure valid.

The covariance matrix considers correlations between error vectors and thus results is higher informational efficiency compared to the Euclidean Norm nonconformity measure. To assess the efficiency of Mahalanobis measure against Norm measure, we calculate the S-criterion and N-criterion. The predicted region sizes per significance level using the action of Mahalanobis measure is significantly smaller than that of the Norm measure. Additionally,

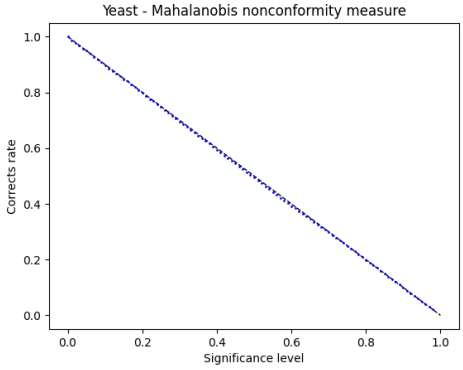


(*a*) Mahalanobis Correct-rate per level ε

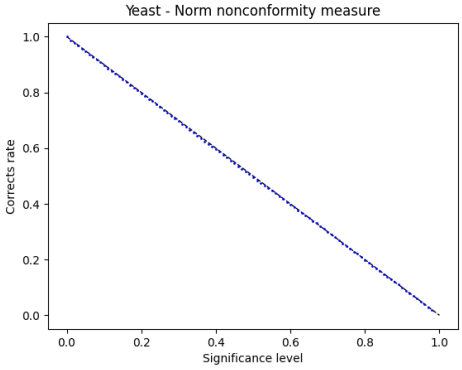


(*b*) Norm Correct-rate per level ε

Figure 2: Mahalanobis and Norm Correct-rate for Emotions dataset.



(*a*) Mahalanobis Correct-rate per level ε



(*b*) Norm Correct-rate per level ε

Figure 3: Mahalanobis and Norm Correct-rate for Yeast dataset.

there is a notable difference in the average sum of the p-values obtained from the two measures.

Our immediate future direction involves developing a method for efficiently generating the power set of labels based on the idea proposed by Maltoudoglou et al. (2022). The size of the power set depends on the number of classes, and by reducing computational complexity, particularly for datasets with numerous classes, we can further explore the application of the Mahalanobis nonconformity measure. Moreover, we plan to examine the formulation of a more informative approach for displaying the predicted region results.

References

Antonis Lambrou and Harris Papadopoulos. Binary relevance multi-label conformal predictor. In *Conformal and Probabilistic Prediction with Applications*, pages 90–104. Springer,

2016.

Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognition*, 122:108271, 2022.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pages 294–306. PMLR, 2022.

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 241–250. Springer, 2014.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002a.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Qualified prediction for large data sets in the case of pattern recognition. In *ICMLA*, pages 159–163, 2002b.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: A tree-based conformal prediction approach. In *Conformal and Probabilistic Prediction with Applications*, pages 488–512. PMLR, 2023.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pages 23–39. Springer, 2016.

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 *Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings 3*, pages 241–250. Springer, 2015.

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, 2013.