Multi-label Conformal Prediction with a Mahalanobis Distance Nonconformity Measure

Kostas Katsios, Harris Papadopoulos

Computational Intelligence Research Lab.
Frederick University

Machine Learning Research Group
Albourne Partners Ltd

The 13th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2024)

September 9, 2024

Presentation Overview

Multi-label classification is a problem category in which each instance can belong to multiple classes simultaneously, resulting in the formation of label-sets.

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 ψ is a subset of C,

$$\psi \subseteq C.$$

Multi-label classification progress studies

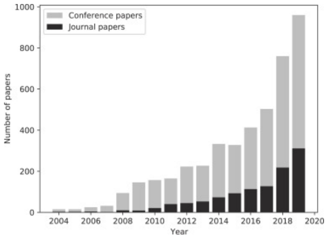


Figure: (Bogatinovski et al. 2022): A summary of the number of papers from the SCOPUS database related to the topic of Multi-label Classification. The vertical axis represents the number of conference and journal papers related to the topic per year.

Paper ([Wang et al. 2017]) published in *Proceedings of the IEEE conference*

"Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases"

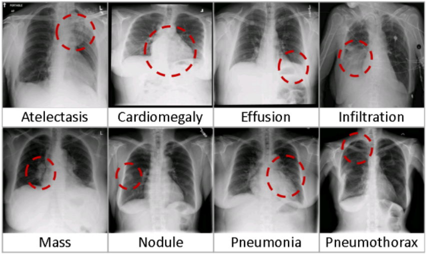


Figure 1. *Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.*

Data: comprises 108,948 frontal-view X-ray images of 32,717 unique patients

Citations: more than 4000

Funding:
- Research Programs of the NIH Clinical Center and National Library of Medicine
- GPU donation by **NVIDIA Corporation**

Multi-label classification techniques fall into two major categories (Tsoumakas and Katakis 2007):

- Algorithm Adaptation (AA) methods:
  Modified versions of multi-class machine learning techniques for predicting sets of labels.
- Problem Transformation (PT) methods:
  Such as:
  - Binary Relevance (BR)
  - Instant Reproduction (IR)
  - Label Power-set (LP)

Differences LP-CP and other multi-label CP methods:

1. Calculation of nonconformity scores and p-values
2. Construction of prediction regions
3. Provided guarantee
4. Computational cost
5. Label dependencies and interactions

Example space symbolism

- Ψ denote a set of label-sets.
- X denote the feature space of which the inputs are represented as vectors of the form,

$$\vec{x}_i = (x_{i_1}, ..., x_{i_s}),$$

  where $X \cong \mathbb{R}^s$ and s is the number of attributes.
- Z denote example space,

$$Z = \{(x_i, \psi_i) \ : \ x_i \in X, \psi_i \in \Psi, i = 1, ..., n\},$$

Training set partitioning

- proper-training set $\{(x_1, \psi_1), ..., (x_q, \psi_q)\}$, where $q \leq n$.
- calibration set $\{(x_{q+1}, \psi_{q+1}), ..., (x_n, \psi_n)\}$.

Nonconformity measure of the calibration instances

$$A : Z \to \mathbb{R} \text{ with } a_i = A\Big(\{(x_1, \psi_1), ..., (x_q, \psi_q)\}, (x_i, \psi_i)\Big), \ i = q + 1, ..., n.$$

Nonconformity measure of the test instances

Let $\mathcal{Y}_j$ denote every assumed label-set for a test instance x_{n+m}.

$$a_{n+m}^{\mathcal{Y}_j} = A\Big(\{(x_1, \psi_1), ..., (x_q, \psi_q)\}, (x_{n+m}, \mathcal{Y}_j)\Big)$$

P-value p of each possible label $\mathcal{Y}_j$

$$p(\mathcal{Y}_j) = \frac{\left| i = q+1, \ldots, n \ : \ a_i \geq a_{n+m}^{\mathcal{Y}_j} \right| + 1}{n - q + 1}$$

Prediction regions for every test instance x_{n+m}

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

We sort the calibration scores in descending order and we denote the ordered calibration scores as a_k^{desc}, for $k = 1, \ldots, n-q$, where $a_1^{desc} < \ldots < a_{n-q}^{desc}$.

Proposition:

For some value ε of the significance level , the minimum integer of which the inequality,

$$\left| \{ i = q+1, \ldots, n \ : \ a_i^{desc} \geq a_{k_\varepsilon}^{desc} \} \right| > \varepsilon(n - q + 1) - 1,$$

holds is,

$$k_\varepsilon = \lfloor \varepsilon(n - q + 1) \rfloor.$$

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

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

Let $\mathcal{P}(C) = \{\mathcal{Y}_j : \mathcal{Y}_j \subseteq C\}$ denote the power-set generated by all combinations of classes.

For every label-set $\mathcal{Y}_j \in \mathcal{P}(C)$, we construct a multi-hot vector $\vec{y}_j = (y_{j_1}, ..., y_{j_c}, ..., y_{j_d})$ as follows,

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

Thus, 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 $\vec{y}_j$.

Notes:

- The empty set in $\mathcal{P}(C)$ corresponds to the zero vector.
- The number of possible multi-hot vectors in Y equals the number 2^d of possible label-sets in $\mathcal{P}(C)$.

Multi-label ICP using Mahalanobis measure

Denote $\vec{o} = \vec{o}(x)$ the predicted probabilities of classifier, for an instance x, where $o \in \mathbb{R}^d$.

We define the linear transformation $r : \mathbb{R}^d \times \{\vec{o}(x)\} \to \mathbb{R}^d$ with,

$$r(\vec{y}, \vec{o}(x)) = |\vec{y} - \vec{o}(x)|.$$

Definition

We define $\vec{r}_i^{\,y_j} = (r_{i_1}, \ldots, r_{i_d})$ as the error vector for instance i related to label-set y_j, such that

$$\vec{r}_i^{\,y_j} = (|y_{j_1} - o_{i_1}|, \ldots, |y_{j_d} - o_{i_d}|),$$

where $\vec{o}_i = (o_{i_1}, \ldots, o_{i_d})$, with $o_{i_k} \in [0, 1]$, $k = 1, \ldots, d$.

Notes:

- The error vectors constitute a subspace R of $\mathbb{R}^d$.
- The linear map r is injective, and thus the label-space Y and the error space R are isomorphic.
- 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.

Let $\vec{y}_j$ denote the true label for calibration instances and assumed label for the test instances.

Euclidean Distance (Norm) nonconformity measure

Maltoudoglou et al. 2022 define a nonconformity measure, for an instance i, using Euclidean Distance as,

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

Mahalanobis Distance nonconformity measure

Definition

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

$$\alpha_i^{y_j} = \sqrt{\left(\vec{r}_i^{y_j}\right)^T \Sigma^{-1} \vec{r}_i^{y_j}}$$

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

Note:

- The covariance matrix takes into account the correlation of the error vectors.
- The Mahalanobis distance is a transformation of the Euclidean distance achieved by using the covariance matrix.
- Σ is symmetric and positive definite.

Algorithm: Multi-label ICP using Mahalanobis measure

Input:

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

Steps:

1. Preprocessing on proper-training data:
   - Calculate the error vectors $\vec{r}_i = |\vec{o}_i - \vec{t}_i|$, $i = 1, \ldots, q$.
   - Form the covariance matrix Σ.
2. Preprocessing on calibration data:
   - Calculate the error vectors $\vec{r}_i = |\vec{o}_i - \vec{t}_i|$, $i = q + 1, \ldots, n$.
   - Calculate the calibration nonconformity scores a_i, $i = q + 1, \ldots, n$, using $\alpha_i^{t_i} = \sqrt{(\vec{r}_i^{\,t_i})^T \ \Sigma^{-1} \ \vec{r}_i^{\,t_i}}$.
   - Sort calibration scores in descending order a_k^{desc}, $k = 1, \ldots, n - q$.
   - Calculate k_ε using $k_\varepsilon = \lfloor \varepsilon(n - q + 1) \rfloor$.
3. Calculate scores $a_{n+m}^{y_j}$, for every possible label-set $\vec{y}_j \in Y$, using $\alpha_i^{y_j} = \sqrt{(\vec{r}_i^{\,y_j})^T \ \Sigma^{-1} \ \vec{r}_i^{\,y_j}}$.

Output:

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

Emotions and Yeast datasets

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

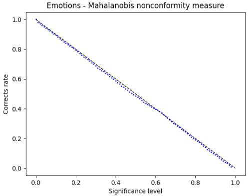
Multi-layer Perceptron (MLP) model

- multiple five fully connected layers
- activation function relu is defined in each layer
- the sigmoid activation function is defined for the probabilistic outputs
- early stopping is set up to avoid overfitting

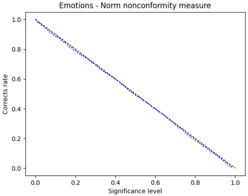
Dataset partitioning

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

Note:
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.

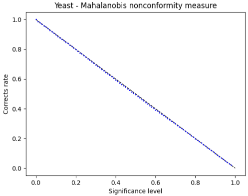


(a) Mahalanobis coverage per level ε

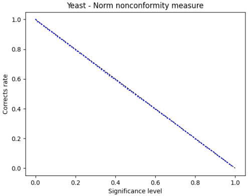


(b) Norm coverage per level ε

Figure 2: Mahalanobis and Norm coverage for Emotions dataset.



(a) Mahalanobis coverage per level ε



(b) Norm coverage per level ε

Figure 3: Mahalanobis and Norm coverage for Yeast dataset.

Table 1: 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

Table 2: 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

Note: The performance results indicate that no substantial classification performance is sacrificed by the use of ICP.

Table: Mahalanobis and Norm S-criterion comparison

	Mahalanobis	Norm
Emotions	547.005	560.869
Yeast	30922.511	81839.323

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

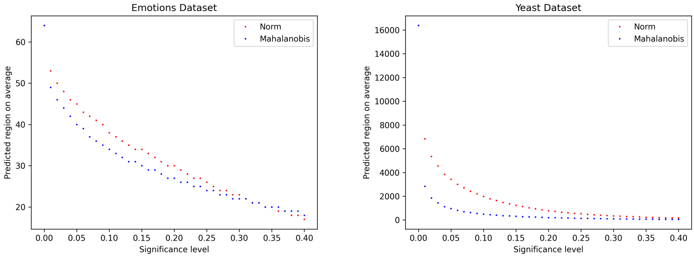


Table: Mean prediction region size as a percentage of the number of possible label-sets

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

Note:
- The number of possible label-sets is 64 and 16.384 for the Emotions and Yeast dataset, respectively.
- In all cases, the Mahalanobis measure produces smaller regions with the values for the Yeast dataset demonstrating an impressive reduction.

Conclusions and future work

Conclusions

- The 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 distance nonconformity measure.
- The prediction region sizes per significance level using the action of Mahalanobis measure is significantly smaller than that of the Norm measure.

Future work

- Formulate the calculation of nonconformity scores based on the nonconformity score of the predicted label-set.
- Develop an approach for efficiently calculating prediction regions (without calculating all p-values)
- Further explore the application of Mahalanobis nonconformity measure.
- Examine the formulation of a more informative ways of presenting the outputs.

References

Bogatinovski, Jasmin et al. (2022). “Comprehensive comparative study of multi-label classification methods”. In: *Expert Systems with Applications* 203, p. 117215.

Lambrou, Antonis and Harris Papadopoulos (2016). “Binary Relevance Multi-label Conformal Predictor”. In: *Conformal and Probabilistic Prediction with Applications*. Springer, pp. 90–104.

Maltoudoglou, Lysimachos et al. (2022). “Well-calibrated confidence measures for multi-label text classification with a large number of labels”. In: *Pattern Recognition* 122, p. 108271.

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

Papadopoulos, Harris (2014). “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*. Springer, pp. 241–250.

Tsoumakas, Grigorios and Ioannis Katakis (2007). “Multi-label classification: An overview”. In: *International Journal of Data Warehousing and Mining (IJDWM)* 3.3, pp. 1–13.

Wang, Xiaosong et al. (2017). “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.