GML_DT: A Novel Graded Multi-label Decision Tree Classifier

Wissal Farsal, Mohammed Ramdani, Samir Anter
Computing Laboratory of Mohammedia (LIM)
FSTM, HassanII University of Casablanca
Morocco

Abstract—**The goal of Graded Multi-label Classification (GMLC) is to assign a degree of membership or relevance of a class label to each data point. As opposed to multi-label classification tasks which can only predict whether a class label is relevant or not. The graded multi-label setting generalizes the multi-label paradigm to allow a prediction on a gradual scale. This is in agreement with practical real-world applications where the labels differ in matter of level relevance. In this paper, we propose a novel decision tree classifier (GML_DT) that is adapted to the graded multi-label setting. It fully models the label dependencies, which sets it apart from the transformation-based approaches in the literature, and increases its performance. Furthermore, our approach yields comprehensive and interpretable rules that efficiently predict all the degrees of memberships of the class labels at once. To demonstrate the model's effectiveness, we tested it on real-world graded multi-label datasets and compared it against a baseline transformation-based decision tree classifier. To assess its predictive performance, we conducted an experimental study with different evaluation metrics from the literature. Analysis of the results shows that our approach has a clear advantage across the utilized performance measures.**

Keywords—Graded multi-label classification; algorithm adaptation; decision tree classifier; label dependencies

I. Introduction

Multi-label classification (MLC) has become an extensively researched and prominent field in machine learning. This is attributed to various real world applications that the traditional task of classification could simply not cover. Instead of predicting one class at a time, MLC predicts multiple classes at the same time. The classes are predicted based on a relevance/non relevance paradigm, while this task has proven to be useful, it remains limited as to the information it provides. Hence, an extension of MLC called Graded Multi-label Classification (GMLC) was proposed in [1].

GMLC assigns a degree of relevance or membership for each label to an instance. The degrees of relevance are gradual memberships in the sense of fuzzy set theory. A Covid-19 article, for example, may belong to three classes {health, economy, society} at the same time. However, the degree of membership to each class differs. The article can fully belong to the class health while it remains somewhat socio-economical.

In this light, all multi-label problems are graded multi-label problems, where the membership degrees are reduced to two binary values, relevant/non relevant. However, the reverse is not true, and reducing the graded multi-label problem to a standard multi-label problem was shown to decrease the predictive performance [1]. Hence the need of graded multi-label classifiers that can generalize the multi-label learning to encompass graded multi-label learning tasks.

Research on multi-label learning in recent years provided solutions in a variety of real-word problems where the traditional learning paradigms were not applicable, ranging from text categorization [2], automatic video and image annotation [3] [4] [5], web mining [6], information retrieval [7] to medical research and bioinformatics [8] [9] [10]. The different algorithms and approaches proposed exploited the transformation and the adaptation methods [11]. The diversity of these approaches is necessary to answer various real-world applications. In bioinformatics, for example, and more specifically genomics, the adaptation of the decision tree classifiers was proven to be very important in deducing comprehensible and readable rules predicting the functional classes of the genes.

Similarly, the ongoing research on Graded multi-label classification aims at developing solutions for real-world problems where the multi-label learning paradigm is not applicable or not optimal. For this purpose, some graded multi-label classifiers were proposed [1] [12] [13] [14]. However, to the best of our knowledge, there are no adaptation-based classifiers for GMLC in the literature.

In this paper, we propose a novel adapted decision tree classifier (GML-DT) that is suited for the graded multi label setting. The main advantages of this approach are its ability to fully model the label dependencies which improves the quality of its predictions. Furthermore, this algorithm is the first adapted tree-based model which makes it the most interpretable existing approach in GML. It is the only model in the literature that constructs a single decision tree from which a set of intelligible and accurate rules can be easily extracted.

The rest of the paper is organized as follows: Section II reviews previous works on GMLC and adapted decision trees in MLC. Section III presents the GML-DT algorithm. Section IV displays the experimental results on real-world graded multi-label data. Finally, Section V concludes this work and introduces future perspectives.

II. Related Work

In this section, we go over related work in both graded multi-label classification and multi-label classification.

A. Graded Multi-label Classifiers

Graded multi-label classification [1] was formalized as an extension of multi-label classification [15] [16], to predict the degrees of relevance of the labels rather than the set of relevant labels. By extension, graded multi-label classifiers fall within two main categories, problem transformation and problem adaptation. The former transforms the multi-label problem into a combination of regular classification tasks for each class label, whereas the later modifies directly the classifier to deal with graded multi-label data.

Cheng et al. [1] proposed a solution by decomposing the problem into an ordinal classification problem and a multi-label classification problem. They introduced two transformation methods, namely the vertical reduction, which predicts the membership degree for each label, and the horizontal reduction which predicts a subset of labels on each grade level. The authors also proceeded to prove the usefulness of graded multi-label classification, by deploying and comparing their approach on both GML data and ML data. Although the model proved to be effective in this setting, it does not model label dependencies. Brinker et al. [12] applied pairwise decomposition using three variants of Calibrated Label Ranking [17], to model the preferences between labels. While these approaches outperformed the predictive model developed in [1], they can only model pairwise dependencies. Lastra et al. [13] proposed a non-deterministic learner based on binary relevance that returns an interval whenever the classification is uncertain for a label. This method relies on a tradeoff between the size of the interval and the improvement of the accuracy.

Laghmari et al. [14] introduced an approach for learning label dependencies and label preferences. This is achieved by using the horizontal decomposition to reduce the problem into a combination of multi-label learning tasks, and then combining pairwise comparisons and classifier chains [18], which is an extension of binary relevance consisting of adding the labels as descriptive attributes.

While transformation methods can be easily implemented with the existing algorithms, their inability to fully model label dependencies and their run time can render them inefficient. This is especially true in cases where an interpretable model is needed, specifically a tree-based one capable of inferring accurate rules, which is the focus of this article. In fact, if one is interested in a model that produces rules identifying the features relevant for the prediction, these approaches would be inefficient and even inapplicable. The approaches in [1] can only identify the features relevant for one class label. Pairwise comparisons are not sufficient to fully model label dependencies. Classifier chains method is not applicable in this setting since it includes the labels as features, which would result in unintelligible rules. Furthermore, these approaches have to build a number of learners, proportional to the number of class labels, which affects their run time and interpretability.

B. Multi-label Decision Tree Classifiers

Clare et al. [19] adapted the c4.5 algorithm to handle multi-label data. The authors modified the formula of the entropy to account for the existence and non-existence of each label, and thus producing a decision tree capable of predicting all the class labels at once. The multi-label entropy is calculated as follows:

$$Entropy(S) = -\sum_{i=1}^{N} p(c_i)\log p(c_i) + q(c_i)\log q(c_i) \quad (1)$$

Where $p(c_i)$ is the probability of the class label c_i

$q(c_i) = 1 - p(c_i)$

Blockeel et al. [20] proposed a hierarchical multi-label decision tree, based on predictive clustering trees [21]. The tree is built by recursively partitioning the data into smaller clusters. This is achieved by finding the best attribute-value that reduces the intra-cluster variance. Where the variance is calculated based on the weighted Euclidean distance. Following this work, Vens et al. [22] presented an empirical study confirming the findings in [19] [20] and thus proving the ineffectiveness of transformation-based decision tree learners in comparison to adapted multi-label decision trees.

III. Graded Multi-Label Decision Tree

A. Formal Task Description

In graded multi-label classification, we have a number of training examples from which we build a classifier that assigns a grade or membership degree to each class label. An instance is represented as a vector x of d attribute values $x = [x_1, \dots, x_d]$ drawn for an input domain $A_1 \times \dots \times A_d$. Given $L = \{\lambda_1, \dots, \lambda_n\}$ a finite set of predefined labels and $M = \{\mu_1, \dots, \mu_m\}$ a finite set of predefined ordered membership degrees such that $\mu_1 < \mu_2 < \cdots < \mu_m$ ranging from complete irrelevance to full relevance. An instance x is assigned a vector of membership degrees $y_x = [y_x^1, \dots, y_x^n]$, where y_x^i corresponds to the degree of relevance of the ith label λ_i for the instance x.

We define a graded multi-label classifier H: $A_1 \times \dots \times A_d \to M^n$ as $(x) = \hat{y}_x$, where $\hat{y}_x = [\hat{y}_x^1, \dots, \hat{y}_x^n]$ corresponds to the set of predicted membership degrees for each label $\lambda_i \in$ L and an instance x.

B. GML_DT: Graded Multi-label Decision Tree

To deal with graded multi-label learning tasks, we propose a novel graded multi-label decision tree algorithm (GML-DT), capable of predicting the membership degrees of all target labels simultaneously.

The GML_DT, given in Algorithm 1, is a greedy model that follows a top-down induction approach for building decision trees. The algorithm takes as input the training set. It starts by searching for the best attribute-value test for the root node. It proceeds to splitting the training set based on the selected test into two partitions, one for which the test succeeds and one for which the test fails, and then calls itself recursively on each partition to construct the left and right subtrees.

Algorithm 1 GML_DT

```
Input: an attribute-valued training set S
 If stopping criterion is True then terminate
 End if
 For each attribute A do
 For each split value v do
  Compute overall entropy for splitting on (A, v)
 End for
 End for
  $(A, v)_{best}$ = Best attribute-value that reduces the overall entropy
 Create a node in the Tree with the best test $(A, v)_{best}$
  $S_1, S_2$ = Induced sub-datasets from S based on the test $(A, v)_{best}$
$Sub\_Tree_1$ = GML_DT($S_1$)
$Sub\_Tree_2$ = GML_DT($S_2$)
   Add $Sub\_Tree_1$ and $Sub\_Tree_2$ to the corresponding branches of
   the Tree
Output: Tree
```

The best attribute-value test is selected by considering all possible split values for each attribute. If the attribute is categorical, the algorithm constructs a test of the form $a_i = v_j$, if it is continuous the test takes the form $a_i \leq v_j$. For each node, the algorithm computes the heuristic values of all the possible attribute-value tests. The heuristic calculates the overall entropy induced by splitting the node on an attribute-value test. The overall entropy as defined in equation (2) is the sum of the weighted entropy of the two partitions created by the split according to their size.

The algorithm then selects the test that reduces this heuristic to put in an internal node. It splits the instances based on the test into two partitions and constructs the subtrees as explained above.

$$Overall\_Entopy = \sum_{i \in \{1,2\}} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

$$Entropy(S_i) = \frac{1}{n} \sum_{j=1}^{n} Entropy(S_i, \lambda_j) \quad (3)$$

Where $Entropy(S_i, \lambda_j)$ is defined as follows:

$$Entropy(S_i, \lambda_j) = -\sum_{k=1}^{m} p(\mu_k) \log p(\mu_k) \quad (4)$$

We modified the formula of the entropy in order to handle graded multi-label classification tasks. We propose a graded multi-label entropy that is computed as the averaged sum of the entropies of class labels. This definition ensures that instances with similar degrees of relevance to the set of labels go in the same subset and thus allowing the prediction of a set of membership degrees for the set of labels in the leaves. We use the majority vote on each class label, to predict its corresponding relevance degree.

The algorithm builds the tree until a stopping criterion is triggered. The stopping conditions are:

- The partition is pure, meaning that all instances have the same degree of relevance for each class label.
- The number of instances in the node are less than a predefined threshold.
- The tree reaches a maximum depth.

C. A Toy Example

To demonstrate the process of building a graded multi-label decision tree, we use a toy example. Table I displays 10 samples of the toy dataset, which originally contains a total of 35 instances described with 3 attributes. a_1 is a categorical feature, a_2 and a_3 are continuous features. The set of class labels is constituted by $\{c_1, c_2, c_3\}$ and the set of degrees is $\{0, 1, 2, 3\}$. This set is equivalent to a set of descriptive nominal counterpart for each degree {not at all, somewhat, almost, fully} characterizing the ordinal levels of relevance of the class labels.

TABLE I. GRADE MULTI-LABEL TOY DATASET

Instances	a_1	a_2	a_3	c_1	c_2	c_3
x_1	A	63	7	3	0	0
x_2	C	29	5	3	3	1
x_3	A	69	13	1	2	3
x_4	B	49	11	1	0	2
x_5	C	51	7	2	1	0
x_6	B	61	5	1	0	2
x_7	C	43	17	2	1	0
x_8	A	69	9	3	0	0
x_9	C	27	13	3	3	1
x_{10}	B	10	14	1	0	2

First, we find the best split by iterating over the attribute columns to get potential split values. The potential split values of a continuous attribute column being the middle values between each consecutive values of the attribute. Then we calculate the overall entropy induced by splitting on an attribute a and a potential split value v.

For example, according the samples in table 1, the potential splits for the attribute a_1 are $(a_1 = A)$, $(a_1 = B)$ and $(a_1 = C)$

Based on these samples, the overall entropy induced by the test $(a_1 = A)$ is calculated as follows:

By applying Equations (2) and (3), we obtain:

$$Overall\_Entropy = \frac{3}{10} \times Entropy(S_1) + \frac{7}{10} \times Entropy(S_2)$$

$$Entropy(S_1) = \frac{1}{3} \times [Entropy(S_1, c_1) + Entropy(S_1, c_2) + Entropy(S_1, c_3)]$$

Where S_1 is the set of instances for which $a_1 = A$, and S_2 is the set of instances for which $a_1 \neq A$.

By applying Equation (4), the entropy of the subset S_1 for the class label c_1 is calculated as follows:

$$Entropy(S_1, c_1) = -p(0) \times \log p(0) - p(1) \times \log p(1) - p(2) \times \log p(2) - p(3) \times \log p(3)$$

$$= 0 - \frac{1}{3} \times \log\frac{1}{3} - 0 - \frac{2}{3} \times \log\frac{2}{3}$$

$$= 0.918$$

After computing $Entropy(S_1, c_2)$ and $Entropy(S_1, c_3)$ following the same process, we get the entropy of the subset S_1:

$$Entropy(S_1) = 0.918$$

In the same way, we calculate the entropy of the subset S_2 and we obtain:

$$Entropy(S_2) = 1.556$$

The overall entropy for splitting on the test $(a_1 = A)$ is 1.364

The overall entropies for the remaining potential splits of the attributes a_1, a_2 and a_3 can be computed following the same process. The potential tests for the continuous features are determined by considering the middle values between each consecutive values.

Fig. 1 displays the decision tree built by GML_DT based on the toy dataset (35 instances). We can infer the five following rules from this decision tree:

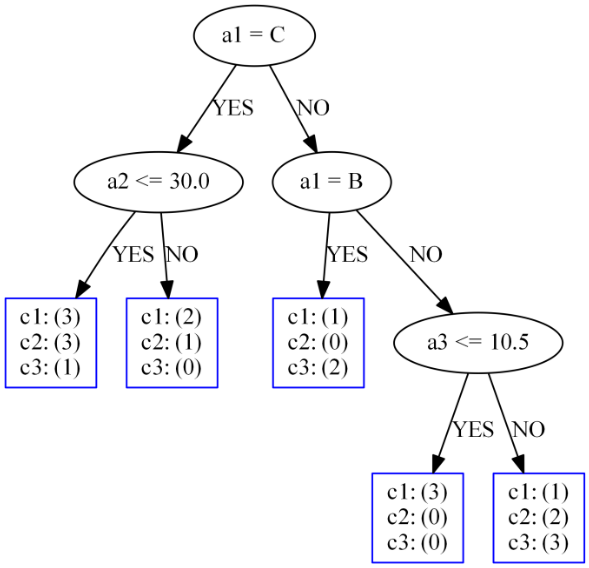

Fig. 1. A Graded Multi-label Decision Tree Constructed by GML_DT.

R_1: IF $a_1 = C$ AND $a_2 \leq 30$ THEN <$Degree(c_1) = 3$, $Degree(c_2) = 3$, $Degree(c_3) = 1$ >

R_2: IF $a_1 = C$ AND $a_2 > 30$ THEN <$Degree(c_1) = 2$, $Degree(c_2) = 1$, $Degree(c_3) = 0$ >

R_3: IF $a_1 = B$ THEN <$Degree(c_1) = 1$, $Degree(c_2) = 0$, $Degree(c_3) = 2$ >

R_4: IF $a_1 \notin \{C, B\}$ AND $a_3 \leq 10.5$ THEN <$Degree(c_1) = 3$, $Degree(c_2) = 0$, $Degree(c_3) = 0$ >

In this toy dataset the attribute a_1 has three possible values A, B or C. Hence, the condition $a_1 \notin \{C, B\}$ is equivalent to $a_1 = A$. Therefore, the fourth rule becomes:

R_4: IF $a_1 = A$ AND $a_3 \leq 10.5$ THEN <$Degree(c_1) = 3$, $Degree(c_2) = 0$, $Degree(c_3) = 0$ >

R_5: IF $a_1 = A$ AND $a_3 > 10.5$ THEN <$Degree(c_1) = 1$, $Degree(c_2) = 2$, $Degree(c_3) = 3$ >

As demonstrated above, the new developed graded multi-label model yields rules that are intelligible and interpretable.

IV. Experimental Study

We conducted an experimental study on real-world datasets, comparing our approach with binary relevance (BR) applied with a state of the art decision tree classifier, under the evaluation metrics from the literature [12] [1]. Cheng et al. [1] generalized some of the common loss functions used in multi-label classification for the graded multi-label setting and introduced the benchmark dataset BeLaE. These performance metrics were then used in the experimental study in [12], which compares the previous work in [1] and three new implemented approaches. This study was carried out on the benchmark dataset BeLaE and two additional real-world graded multi-label datasets curated by the authors, which makes it the most extensive work on GMLC compared to the rest of the work in the literature. For the purpose of conformity, we used the same datasets in our experimental study, along with three of the evaluation metrics from this previous work.

The experimental study is obtained by carrying out 10-fold cross validation on each single dataset. The same folds were used for both experiments on GML_DT and a baseline decision tree classifier applied with binary relevance approach (BR_DT).

We developed the GML_DT algorithm proposed in this paper from scratch using Python. The baseline BR_DT is implemented using the Scikit-learn library [23].

A. Evaluation Metrics

We evaluated the predictive performance of the algorithm based on three metrics; the hamming loss, which corresponds to the mean deviation of the predicted membership degrees to the true membership grades:

$$E_H(\hat{y}_x, y_x) = \frac{\sum_{i=1}^{n} AE(\hat{y}_x^i, y_x^i)}{(m-1)n} \qquad (5)$$

Where AE is the absolute error of the predicted membership degree and it is defined as:

$$AE: M \times M \to L, AE(\mu_i, \mu_j) = |i - j|$$

The vertical 0-1 loss measures the percentage of class labels with incorrectly predicted degrees of relevance:

$$E_{0/1}(\hat{y}_x, y_x) = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_x^i \neq y_x^i) \qquad (6)$$

Where I is the indicator function.

The C-index measures the pairwise ranking errors between the true membership set and the predicted membership set.

$$E_{CI} = \frac{\sum_{i<j} \sum_{(\lambda,\lambda') \in M_i \times M_j} S(h_x(\lambda), h_x(\lambda'))}{\sum_{i<j} |M_i| \times |M_j|} \qquad (7)$$

Where $M_i = \{\lambda \in L | L_x(\lambda) = \mu_i\}$

$L_x(\lambda)$ is a function returning the degree of membership of the label λ for an instance x.

$h_x(\lambda)$ is the predicted degree of membership of the label λ for an instance x.

and $S(u,v) = I(u > v) + \frac{1}{2}\, I(u = v)$

B. Datasets

The datasets used for the experimental study are the BeLa-E which is a benchmark dataset introduced in [1] consisting of 100 variants, 50 datasets predicting 5 labels and 50 datasets predicting 10 labels. BeLaE was constructed based on a survey conducted with 1930 students, in order to grade on a finite ordinal scale the importance of different job properties.

Movies [12] is a dataset collected from 1967 movies where each movie is graded on its level of membership to five descriptive categories, e.g. its level of humor, action, suspense.

Medical [12], based on 1953 radiology reports, annotated with a set of ICD-9-CM disease/diagnosis classification codes. This dataset was adapted from the multi-label dataset by taking into account the level of agreement of the annotators.

Table II summarized the different properties of the aforementioned datasets.

C. Results

Table III displays the experimental results of GML_DT in comparison to BR_DT. We averaged the evaluations across the 10-fold cross validation for each single dataset. For the benchmark datasets, BeLaE (n=5) and BeLaE (n=10), we averaged the performance over the 50 variants of each. We summarize their predictive measures in terms of the mean and the standard deviation.

The experimental study conducted reaches the following conclusions:

GML-DT outperforms the baseline classifier BR_DT in terms of predictive performance according to all three evaluation metrics used in this experiment for all four datasets.

The performance metrics used in these experiments evaluate the results along three dimensions depicting the disparity, accuracy and pairwise ranking errors between the true membership set and the predicted membership set. Hence, GML_DT yields more accurate rules across all these different dimensions.

TABLE II. OVERVIEW OF THE GRADED MULTI-LABEL DATASETS, AND THEIR PROPERTIES: NUMBER OF INSTANCES, NUMBER OF ATTRIBUTES, NUMBER OF CLASS LABELS AND THE NUMBER OF GRADES

Datasets	**Instances**	**Attributes**	**Labels**	**Grades**
BeLaE n=5	1930	45	5	5
BeLaE n=10	1930	40	10	5
Movies	1967	27002	5	4
Medical	1953	1602	204	4

TABLE III. EXPERIMENTAL RESULTS FOR EACH DATASET ACCORDING TO THE HAMMING LOSS, THE VERTICAL 0/1 LOSS AND THE C-INDEX

Datasets	**Evaluation Measures**	**GML_DT**	**BR_DT**
BeLa-E n=5	Hamming Loss	0.168 ±0.018	0.257 ±0.026
	Vertical 0-1 Loss	0.526 ±0.039	0.689 ±0.032
	C-Index	0.264 ±0.047	0.374 ±0.055
BeLa-E n=10	Hamming Loss	0.174 ±0.011	0.259 ±0.017
	Vertical 0-1 Loss	0.540 ±0.020	0.690 ±0.023
	C-Index	0.263 ±0.030	0.361 ±0.040
Movies	Hamming Loss	0.172	0.253
	Vertical 0-1 Loss	0.424	0.536
	C-Index	0.247	0.368
Medical	Hamming Loss	0.002	0.010
	Vertical 0-1 Loss	0.006	0.017
	C-Index	0.135	0.448

Furthermore, GML_DT has a smaller model size compared to BR_DT. In fact, GML_DT is a single model that predicts the membership degrees relative to the set of class labels simultaneously. It builds a single decision tree that identifies the attribute-value conditions relevant for the prediction of the complete set of degrees associated to the label set. On the other hand, BR_DT runs $|L|$ times, and results in $|L|$ constructed decision trees, one for each class label, which not only affects its execution time and complexity but also its interpretability. The higher the number of labels, the more complicated the model gets and therefore the less effective it becomes for retrieving useful and comprehensible rules.

Moreover, if we were to compare GML_DT to the state of the art approaches solely based on the results reported in [1] and [12], we can deduce that GML_DT outperforms the model in [1] across all three metrics. Furthermore, it has better results for the hamming loss and the vertical 0-1 loss compared to the full CLR and Joined CLR while it remains very competitive against the Horizontal CLR [12].

V. CONCLUSION

We present a graded multi-label decision tree classifier, GML_DT, which generalizes the multi-label setting by predicting the membership degrees of the target labels instead of the binary relevance/non relevance. This approach utilizes the interpretability of decision tree classifiers and produces readable and comprehensive trees, which can be translated into useful, homogenous rules. GML_DT is also the first adaptation algorithm in the literature that fully models label dependencies, resulting in an increase of the predictive performance and the quality of the deduced rules.

The proposed algorithm is based on a new adapted graded multi-label heuristic that allows the algorithm to split based on the homogeneity of the combined set of labels, and ultimately retuning a vector containing the majority grade in each class label. We carried out an experimental study on real-world graded multi-label datasets, and evaluated our approach against a state of the art transformation-based decision tree classifier.

Our model is more interpretable and has the best predictive quality according to a variety of performance measures from the GMLC literature.

This paper constitutes a preliminary presentation of GML_DT, we are currently investigating further adaptations of the heuristic to the GML setting in order to improve the predictive performance of the model. Moreover, we are working on reducing the complexity of the generated tree via an adapted post pruning method.

References

[1] CHENG, Weiwei, DEMBCZYNSKI, Krzysztof, et HÜLLERMEIER, Eyke. Graded multilabel classification: The ordinal case. In : ICML. 2010.

[2] CHEN, Guibin, YE, Deheng, XING, Zhenchang, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In : 2017 international joint conference on neural networks (IJCNN). IEEE, 2017. p. 2377-2383.

[3] LIU, Yang, WEN, Kaiwen, GAO, Quanxue, et al. SVM based multi-label learning with missing labels for image annotation. Pattern Recognition, 2018, vol. 78, p. 307-317.

[4] ZHU, Feng, LI, Hongsheng, OUYANG, Wanli, et al. Learning spatial regularization with image-level supervisions for multi-label image classification. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 5513-5522.

[5] MARKATOPOULOU, Foteini, MEZARIS, Vasileios, et PATRAS, Ioannis. Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. IEEE transactions on circuits and systems for video technology, 2018, vol. 29, no 6, p. 1631-1644.

[6] PRABHU, Yashoteja, KAG, Anil, GOPINATH, Shilpa, et al. Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation. In : Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018. p. 441-449.

[7] ZHANG, Zheng, ZOU, Qin, LIN, Yuewei, et al. Improved deep hashing with soft pairwise similarity for multi-label image retrieval. IEEE Transactions on Multimedia, 2019, vol. 22, no 2, p. 540-553.

[8] BUSTOS, Aurelia, PERTUSA, Antonio, SALINAS, Jose-Maria, et al. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 2020, vol. 66, p. 101797.

[9] ZHOU, Jian-Peng, CHEN, Lei, et GUO, Zi-Han. iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. Bioinformatics, 2020, vol. 36, no 5, p. 1391-1396.

[10] ZHANG, Jingpu, ZHANG, Zuping, WANG, Zixiang, et al. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. Bioinformatics, 2018, vol. 34, no 10, p. 1750-1757.

[11] ZHOU, Zhi-Hua et ZHANG, Min-Ling. Multi-label Learning. 2017.

[12] BRINKER, Christian, MENCÍA, Eneldo Loza, et FÜRNKRANZ, Johannes. Graded multilabel classification by pairwise comparisons. In : 2014 IEEE International Conference on Data Mining. IEEE, 2014. p. 731-736.

[13] LASTRA, Gerardo, LUACES, Oscar, et BAHAMONDE, Antonio. Interval prediction for graded multi-label classification. Pattern Recognition Letters, 2014, vol. 49, p. 171-176.

[14] LAGHMARI, Khalil, MARSALA, Christophe, et RAMDANI, Mohammed. Learning Label Dependency and Label Preference Relations in Graded Multi-label Classification. Computational Intelligence for Pattern Recognition, 2018, p. 115-164.

[15] TSOUMAKAS, Grigorios, KATAKIS, Ioannis, et VLAHAVAS, Ioannis. Mining multi-label data. In : Data mining and knowledge discovery handbook. Springer, Boston, MA, 2009. p. 667-685.

[16] TSOUMAKAS, Grigorios et KATAKIS, Ioannis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 2007, vol. 3, no 3, p. 1-13.

[17] FÜRNKRANZ, Johannes, HÜLLERMEIER, Eyke, MENCÍA, Eneldo Loza, et al. Multilabel classification via calibrated label ranking. Machine learning, 2008, vol. 73, no 2, p. 133-153.

[18] READ, Jesse, PFAHRINGER, Bernhard, HOLMES, Geoff, et al. Classifier chains for multi-label classification. Machine learning, 2011, vol. 85, no 3, p. 333-359.

[19] CLARE, Amanda et KING, Ross D. Knowledge discovery in multi-label phenotype data. In : European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2001. p. 42-53.

[20] BLOCKEEL, Hendrik, SCHIETGAT, Leander, STRUYF, Jan, et al. Decision trees for hierarchical multilabel classification: A case study in functional genomics. In : European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 2006. p. 18-29.

[21] BLOCKEEL, Hendrik, DE RAEDT, Luc, et RAMON, Jan. Top-down induction of clustering trees. arXiv preprint cs/0011032, 2000.

[22] VENS, Celine, STRUYF, Jan, SCHIETGAT, Leander, et al. Decision trees for hierarchical multi-label classification. Machine learning, 2008, vol. 73, no 2, p. 185.

[23] PEDREGOSA, Fabian, VAROQUAUX, Gaël, GRAMFORT, Alexandre, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011, vol. 12, p. 2825-2830.