Marek Wydmuch

Addressing the long-tail problem in extreme multi-label classification

Doctoral Dissertation

Submitted to the Discipline Council
of Information and Communication
Technology
of Poznan University of Technology

Advisor: Krzysztof Dembczyński, Eng. Ph.D., Habil.

Poznan · 2025

Marek Wydmuch

Problem długiego ogona w ekstremalnej klasyfikacji wieloetykietowej

Rozprawa doktorska

Przedłożono Radzie Dyscypliny
Informatyka Techniczna
i Telkomunikacja
Politechniki Poznańskiej

Promotor: Krzysztof Dembczyński, Inż. Dr Hab.

Poznań · 2025

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computing Science.

Marek Wydmuch
Machine Learning Laboratory
Faculty of Computing and Telecommunications
Institute of Computing Science
Poznan University of Technology
mwydmuch@cs.put.poznan.pl

The dissertation was typeset by the author in LaTeX.
The cover and back cover were designed by the author.

This dissertation and associated materials can be downloaded from:
https://www.cs.put.poznan.pl/mwydmuch/dissertation

Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60–965 Poznan, Poland
https://www.cs.put.poznan.pl

Computational experiments were partially performed in Poznan Supercomputing and Networking Center.

Abstract

Extreme multi-label classification (XMLC) is a learning task of assigning multiple labels to instances from an extremely large pool of possible labels, reaching an order of hundreds of thousands or even millions. Such scenarios are increasingly prevalent in modern applications like document and media tagging, recommendation systems, and web advertising. However, XMLC faces significant challenges, notably computational complexity and severe statistical issues arising from the sparsity of labels, commonly referred to as the "long-tail" problem. This dissertation specifically focuses on the "long-tail" labels, which, despite their infrequency, hold significant value in many real-world applications. However, the current XMLC metrics, such as precision@k, inadequately capture performance on these rare labels, allowing classifiers to ignore them without impacting the metric values, motivating the need for an alternative way of evaluating performance on the "long tail". Consequently, this thesis proposes the use of macro-averaged metrics, which average binary metrics calculated for each label, making them all equally important. The common settings in XMLC often require predicting exactly k labels. This constraint, combined with macro-averaged metrics, creates a unique optimization challenge that was not considered before. The core hypothesis examined in this dissertation asserts the existence of consistent inference algorithms for optimizing these macro-averaged metrics budgeted at k. The main contributions include deriving Bayes-optimal classifiers for both classic instance-wise and macro-averaged metrics. The latter is being analyzed under two distinct statistical frameworks – expected test utility (ETU) and population utility (PU). In all cases, regret bounds that establish consistency under realistic conditions are provided, and computationally efficient inference algorithms are proposed. Empirical evaluations validate theoretical results, demonstrating that introduced methods significantly improve results on macro-averaged metrics. At the same time, they can maintain good performance on classical metrics and are computationally efficient, making them suitable for the XMLC problems.

Introduction and outline

Machine learning and extreme multi-label classification

Machine learning is currently a broad and fast-growing sub-field of artificial intelligence that sits between domains of computer science, mathematical optimization, information theory, statistics, and even neuroscience. The fundamental idea behind machine learning is to enable computers to discover algorithms for solving specific tasks that would be too difficult or too costly to be solved by human programmers. Instead of being provided with explicit instructions, a computer is presented with data related to the problem (e.g., examples of correctly performed tasks) and "learns" how to solve it based on that data. Recent advancements in the development of machine learning have led to its being successfully applied to a growing number of applications, which leads to the emergence of new research directions. On a high level, machine learning approaches are traditionally divided into three categories, which correspond to learning paradigms, depending on the nature of the data available to the learning system:

- Supervised learning: A computer is presented with example inputs and their desired outputs (labels), given by a "teacher". We call the set of such examples a training set. The goal is to learn from these a general function that maps inputs to outputs.
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find hidden structure in its inputs.
- Reinforcement learning: A computer program (usually called an agent in this context) interacts with a dynamic environment in which it must achieve a certain goal. As it performs different actions, the environment provides feedback that is analogous to rewards. The goal is to learn a strategy that maximizes the sum of the rewards.

This categorization is rather a simplistic attempt at pigeonholing numerous machine

learning settings and algorithms.

Nevertheless, in this thesis, we will focus directly on the sub-class of supervised learning called classification. Classification involves assigning a discrete label, or a number of them, to an input instance based on its features as correctly as possible (e.g., give names of objects visible in the image, diagnose a disease based on patient results, assign relevant categories and tags to the article, and predict products that have a chance to be bought by a specific customer.) Classification is one of the most widely studied and applied machine learning tasks across numerous domains.

Simple classification tasks typically involve assigning a single label to each instance (multi-class classification) or determining the presence or absence of a single characteristic (binary classification). However, many real-world problems require assigning multiple labels to a single instance simultaneously. This type of classification problem, known as multi-label classification, introduces additional complexity compared to binary or multi-class classification. For example, a news article might belong to multiple categories such as "politics," "economy," and "international affairs"; an image might contain multiple objects like "person," "car," and "building"; a customer might buy many products from the store; or a patient might have multiple medical conditions simultaneously.

Extreme multi-label classification (XMLC) extends the multi-label classification paradigm to settings with extraordinarily large numbers of potential labels, ranging from thousands to millions, with only a small subset of a few being relevant for each input instance. This scenario is increasingly common in modern applications such as:

1. Documents and media tagging for search: Identifying relevant topics, categories, or keywords in documents and media (text, image, sound, video) from a vocabulary of millions of possible tags [Dekel and Shamir, 2010, Deng et al., 2011, Agrawal et al., 2013].
2. Recommendation: Suggesting a small subset of relevant items from a huge inventory to customers or matching them with other items [Weston et al., 2013, Zhuo et al., 2020, Song et al., 2020], or search queries [Medini et al., 2019, Chang et al., 2020].
3. Web advertising: Matching ads to search queries or web pages from a large collection of advertisements [Prabhu and Varma, 2014, Jain et al., 2019].

The scale of these problems makes XMLC different from standard multi-label classification, requiring different techniques and algorithms. In XMLC, both the feature space and labels are high-dimensional, posing several key challenges:

1. Computational efficiency: multi-label classification methods often have linear complexity with the number of labels or higher to account for complex dependencies between labels. Applying these methods becomes impractical when dealing with a large number of labels. Efficient algorithms that can train and predict in sublinear time are essential in this setting.
2. Statistical challenges: with a large number of potential labels, most of them

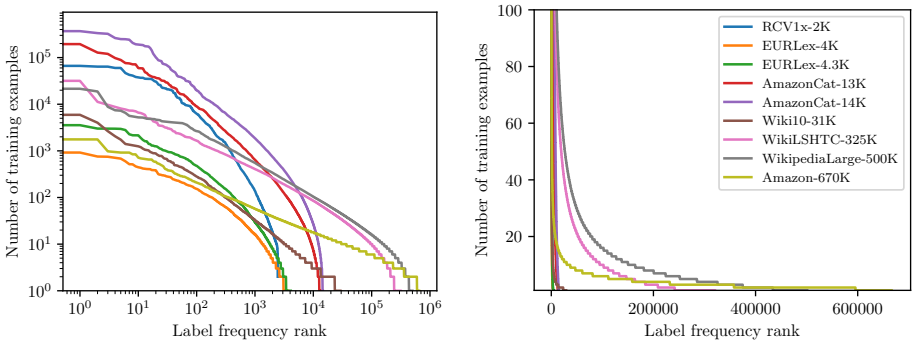

Figure 1: Label frequency in XMLC datasets. The x-axis shows the label rank when sorted by the frequency of positive instances and the y-axis gives the number of the positive instances. The first plot uses a log scale for both the x- and y-axis; the second uses a linear scale but crops the y-axis at 100.

appear in only a tiny fraction of instances, leading to severe data sparsity and imbalance issues. A large pool of labels also makes labeling prone to the nose, especially labels going missing, as it is nearly impossible for annotators to carefully evaluate the whole set of labels when it is so large.

Long-tailed label distribution

While in this thesis, we touch on the aspect of efficient algorithms, we especially focus on the issue of high label imbalance and sparsity. Given the large number of labels in XMLC datasets reaching hundreds of thousands or even millions, and their adherence to Zipf's law [Adamic and Huberman, 2002], it is not surprising that many labels are very sparse, and therefore the label distribution is strongly "long-tailed" [Bhatia et al., 2015, Babbar and Schölkopf, 2017]. In Figure 1, we plot the distribution of label frequencies of nine popular benchmark datasets[1] from the XMLC repository [Bhatia et al., 2016], clearly showing that only a small fraction of labels are well represented in the whole set, while the rest form a long tail with very few examples each.

Paradoxically, in many applications, these rare labels are considered to be more important, as being more informative or more satisfying when predicted correctly, e.g., very general tags are rarely useful for finding information, the user feels better when a good recommendation of an item that was unknown to them instead of most popular one. Posing the critical research question, what is the best way to accommodate these tail labels.

[1]We will use the same set of nine datasets throughout this thesis.

Performance metrics in XMLC

When training a classifier, one usually aims to achieve the highest quality of predictions, expressed in the form of a metric that scores the classifier's output in comparison to the expected output. To evaluate the classifier's generalization capabilities, the additional (test) set, which was not seen by the classifier before, is used. The test set, like the training set, needs to contain both the input and the expected output pairs. The simplest metric in classification is zero-one accuracy, which assigns a score of 1 if the classifier's prediction exactly matches the expected output for a given input and 0 otherwise. The scores are then averaged over all instances in the test set.

Due to the specificity of a setting, the XMLC area adopted its own set of standard metrics for evaluating classifiers. In traditional multi-label classification, the classifier's decision is evaluated either for each label or on the level of the whole positive and negative label sets. Due to the nature of XMLC applications, like search and recommendation, one usually does not care about the precise decision for each label, but rather obtaining a set of labels, often of specific size, that will maximize the evaluation metric. The size of the prediction set can often be indicated by a graphical interface having a specific number of slots for presenting search results, recommendations, or ads [Cremonesi et al., 2010, Chang et al., 2021]. To reflect that, it is popular to evaluate using the so-called "at k" (@k) metrics, for which the classifier returns precisely k labels for an instance (sometimes also with specified order). The most popular metric is precision@k, which counts how many of the k predicted labels were in the set of true labels, followed by metrics like recall@k, or (n)DCG@k.

It has been noticed that, under high label imbalance, competitive performance on these metrics can be achieved by correctly predicting the most popular labels [Jain et al., 2016, Wei and Li, 2020]. Therefore, the community has started to develop new metrics that prefer "rewarding" [Ye et al., 2020], "diverse" [Babbar and Schölkopf, 2019], and "rare and informative" [Prabhu et al., 2018a] labels over the frequently occurring head labels.

Jain et al. [2016] addressed this problem as the first ones by introducing propensity-scored performance metrics, which quickly gained popularity and became a standard in the XMLC community. These metrics are weighted variants of standard metrics like precision@k, that give increased weight to tail labels, but these weights have been derived from the perspective of missing labels, and the interpretation of the results on these metrics is not straightforward.

Motivation

We have noticed that even the popular propensity-scored precision only slightly better distinguishes classifiers that are good at predicting tail labels from the ones

that are not. This observation motivates the search for alternative metrics better suited to evaluating performance in long-tail scenarios.

In multi-label classification, one can consider a wide spectrum of measures that are usually divided into three categories based on the averaging scheme, namely instance-wise, micro-, and macro-averaging. Instance-wise measures are defined, as the name suggests, on the level of a single instance. Typical examples are already mentioned precision@k, recall@k, (n)DCG@k, Hamming loss, and the instance-wise F_β-measure. Micro-averages are defined on a single confusion matrix that accumulates true positives, false positives, false negatives, and true negatives from all the labels. Macro-averages require a binary metric to be applied to each label separately and then averaged over the labels. In general, any binary metric can be applied in any of the above averaging schemes. Not surprisingly, some of the metrics, for example, Hamming loss, lead to the same form of the final metric regardless of the scheme used. One can also consider the wider class of measures that are defined as general aggregation functions of label-wise confusion matrices.

The macro-averaged metric seems to be very attractive in the context of evaluating long-tail performance in XMLC, as they treat all the labels equally important, preventing the labels with a small number of positive examples from being overshadowed. Because of that, we take a closer look at these metrics with a prediction budget of k, which naturally fits the XMLC setting. The budget requires the prediction algorithm to choose labels "wisely," often leading to the presence of long-tail labels in the set of predicted labels. It is also natural in typical applications of XMLC, like recommendation systems, in which the number of slots is dictated by the user interface.

To illustrate the emphasis that macro-averaged put on tail-labels, in Table 1, we compare two classifiers,[2] The first was trained on all labels, and the second was trained only on 20% of the head labels, with the rest being discarded as they would not exist. Then, both models are evaluated using a complete set of labels. The standard metrics are only slightly perturbed by reducing the label space to the head labels. The propensity-scored precision decreases more significantly, sometimes even to 35%, but this drop in performance does not fully reflect the inability of a classifier to predict 80% of labels. In contrast, macro-averaged metrics (precision, recall, and F_1-measure) with a budget at k, as well as coverage@k, a popular auxiliary measure in XMLC [Jain et al., 2016, Babbar and Schölkopf, 2019, Wei and Li, 2020, Schultheis et al., 2022], decrease much more drastically, even over 70%, much better reflecting the classifier's inability to predict the remaining 80% of labels. These results show that budgeted-at-k macro measures might be very attractive in the context of long tails.

[2]We use an ensemble of probabilistic label trees [Jasinska-Kobus et al., 2020], the same method we use later in Chapter 9 for other experiments.

Table 1: Results (%) of a classifier trained on the full set of labels and a classifier trained with only 20% of head labels (most frequent labels) on different metrics budgeted at k (@k). The difference smaller than 5% is colored green, smaller than 30% orange, and bigger using red.

Metric	With all labels			With top 20% labels					
	@1	@3	@5	@1	(diff.)	@3	(diff.)	@5	(diff.)
RCV1x-2K									
Precision	89.99	72.18	51.67	89.96	(-0.03%)	72.08	(-0.14%)	51.57	(-0.21%)
Propensity-scored Prec.	96.94	78.69	56.92	96.84	(-0.11%)	78.36	(-0.42%)	56.52	(-0.72%)
Recall	40.31	74.60	81.17	40.29	(-0.06%)	74.49	(-0.14%)	81.02	(-0.19%)
nDCG	89.99	75.87	60.65	89.96	(-0.03%)	75.79	(-0.11%)	60.56	(-0.15%)
Macro-Precision	10.04	13.79	13.37	8.93	(-11.05%)	9.26	(-32.81%)	7.37	(-44.86%)
Macro-Recall	1.24	4.65	7.61	1.16	(-6.32%)	3.81	(-18.16%)	5.66	(-25.68%)
Macro-F_1	1.74	5.44	7.62	1.61	(-7.45%)	4.28	(-21.38%)	5.34	(-29.89%)
Coverage	12.42	26.14	35.91	11.24	(-9.51%)	17.18	(-34.27%)	18.89	(-47.39%)
EURLex-4.3K									
Precision	91.40	81.21	68.50	90.55	(-0.93%)	79.30	(-2.35%)	65.88	(-3.83%)
Propensity-scored Prec.	151.59	139.26	120.68	144.54	(-4.65%)	127.92	(-8.14%)	107.07	(-11.28%)
Recall	20.81	53.30	71.61	20.59	(-1.07%)	51.88	(-2.66%)	68.68	(-4.09%)
nDCG	91.40	83.62	74.24	90.55	(-0.93%)	81.95	(-2.00%)	72.01	(-3.01%)
Macro-Precision	14.39	22.55	25.36	11.84	(-17.70%)	13.83	(-38.69%)	12.22	(-51.83%)
Macro-Recall	4.45	14.15	22.57	3.26	(-26.71%)	9.04	(-36.10%)	12.97	(-42.53%)
Macro-F_1	6.16	16.41	22.73	4.69	(-23.99%)	10.40	(-36.61%)	12.20	(-46.33%)
Coverage	15.62	27.09	34.54	13.18	(-15.59%)	18.31	(-32.41%)	19.46	(-43.66%)
AmazonCat-14K									
Precision	89.28	69.10	54.63	89.25	(-0.03%)	68.77	(-0.47%)	52.99	(-3.00%)
Propensity-scored Prec.	95.23	77.83	64.44	94.32	(-0.96%)	75.77	(-2.65%)	59.86	(-7.10%)
Recall	39.40	69.11	83.63	39.40	(-0.01%)	68.84	(-0.39%)	81.73	(-2.26%)
nDCG	89.28	73.43	62.34	89.25	(-0.03%)	73.18	(-0.34%)	61.14	(-1.91%)
Macro-Precision	25.19	39.74	41.82	12.81	(-49.15%)	13.41	(-66.26%)	10.94	(-73.83%)
Macro-Recall	2.74	15.50	36.67	1.32	(-51.87%)	6.83	(-55.91%)	12.73	(-65.27%)
Macro-F_1	4.53	20.09	36.57	2.18	(-51.97%)	8.29	(-58.75%)	11.39	(-68.87%)
Coverage	30.25	54.89	69.59	15.57	(-48.52%)	18.91	(-65.56%)	19.87	(-71.45%)
WikiLSHTC-325K									
Precision	63.44	41.94	31.14	59.32	(-6.49%)	37.75	(-9.98%)	27.57	(-11.45%)
Propensity-scored Prec.	163.80	118.44	90.78	133.40	(-18.56%)	89.06	(-24.80%)	66.02	(-27.27%)
Recall	28.11	47.00	54.58	25.39	(-9.68%)	40.54	(-13.74%)	46.18	(-15.40%)
nDCG	63.44	46.72	37.94	59.32	(-6.49%)	42.51	(-9.00%)	34.16	(-9.95%)
Macro-Precision	12.83	18.61	18.73	7.10	(-44.64%)	7.25	(-61.06%)	6.03	(-67.82%)
Macro-Recall	6.53	15.71	20.72	3.57	(-45.42%)	7.38	(-53.00%)	9.17	(-55.75%)
Macro-F_1	7.78	15.19	17.30	4.16	(-46.61%)	6.33	(-58.34%)	6.31	(-63.54%)
Coverage	16.61	29.71	35.53	10.80	(-34.97%)	15.64	(-47.38%)	17.10	(-51.87%)
WikipediaLarge-500K									
Precision	66.83	47.79	37.41	60.88	(-8.91%)	42.00	(-12.11%)	32.46	(-13.22%)
Propensity-scored Prec.	187.95	142.69	113.72	139.25	(-25.91%)	98.25	(-31.15%)	76.40	(-32.82%)
Recall	21.86	39.77	48.07	18.75	(-14.20%)	32.43	(-18.45%)	38.55	(-19.81%)
nDCG	66.83	52.07	43.72	60.88	(-8.91%)	46.20	(-11.26%)	38.45	(-12.05%)
Macro-Precision	13.09	19.92	21.04	5.98	(-54.31%)	6.39	(-67.93%)	5.74	(-72.71%)
Macro-Recall	6.64	16.16	21.83	2.77	(-58.36%)	5.81	(-64.08%)	7.47	(-65.80%)
Macro-F_1	7.84	15.94	18.99	3.23	(-58.76%)	5.15	(-67.67%)	5.51	(-70.99%)
Coverage	16.58	30.83	37.87	9.31	(-43.89%)	13.92	(-54.85%)	15.66	(-58.64%)
Amazon-670K									
Precision	44.97	40.20	36.71	41.01	(-8.80%)	34.38	(-14.48%)	28.95	(-21.13%)
Propensity-scored Prec.	288.28	273.84	259.09	229.95	(-20.23%)	197.65	(-27.82%)	167.45	(-35.37%)
Recall	9.35	23.31	34.39	8.00	(-14.47%)	19.05	(-18.27%)	26.19	(-23.83%)
nDCG	44.97	41.29	38.58	41.01	(-8.80%)	35.91	(-13.03%)	31.75	(-17.70%)
Macro-Precision	4.78	10.46	14.21	3.38	(-29.24%)	5.31	(-49.20%)	5.43	(-61.77%)
Macro-Recall	3.08	9.04	14.41	2.04	(-33.75%)	5.20	(-42.51%)	7.23	(-49.82%)
Macro-F_1	3.42	9.01	13.39	2.26	(-33.74%)	4.70	(-47.82%)	5.59	(-58.23%)
Coverage	5.82	13.78	19.79	4.64	(-20.33%)	9.02	(-34.51%)	11.01	(-44.37%)

In this thesis, we examine the long-tail label problem through the lens of evaluation metrics. We propose the usage of macro-averaged metrics budgeted at k as an alternative, focusing on "tail-labels" performance and being easy to interpret and investigate the problem of optimizing them. Interestingly, optimization of the macro-averaged metrics can be considered under two distinct statistical frameworks – expected test utility (ETU) (also known in the literature as the decision theoretic approach (DTA) and population utility (PU) (also known as empirical utility maximization (EUM) [Ye et al., 2012, Dembczyński et al., 2017]. Under the ETU framework, the goal is to optimize the expected value of a metric on a given test set. While PU aims to optimize a metric value calculated on the expected value of a confusion matrix over the population. While the macro-averaged metrics, as well as a more general class of metrics defined on the label-wise confusion matrices, have been well-studied in multi-label classification under both frameworks, the budgeted @k variants have not, to the best of our knowledge. Furthermore, the previously introduced optimization frameworks cannot be directly applied in the budgeted setup. Since the optimization problems for different labels are tightly coupled with each other through the budget of k predictions, the final problem is much more difficult.

One of the first results of this work is an observation that there exists a class of label-wise metrics budgeted at k, that can be linearly decomposed over labels into binary utilities, which includes both standard instance-wise weighted metrics, like precision@k, and macro-averaged metrics, possibly allowing to obtain general results under a unified framework. Because of that, we are interested in investigating whether consistent inference algorithms exist to optimize label-wise metrics budgeted at k and if they are compatible with existing classifiers for XMLC. Due to extremely large label space, almost all existing XMLC methods aim to provide top k candidates with scores assigned to individual labels. Preferably, these scores reflect (or can be calibrated to reflect) estimates of the label marginal conditional probabilities. To be practical, an algorithm for optimizing label-wise metrics budgeted at k should operate over the marginal conditional probabilities of labels, allowing it to work on top of existing and future methods by being agnostic to their architecture. Additionally, we would like such an algorithm to be consistent, meaning that with the growing size of the training set, the regret of the classifier decreases to zero.

Aim and scope

Given the motivations presented above, the hypothesis of this dissertation is formulated as follows:

> **There exist consistent inference algorithms for optimizing label-wise metrics budgeted at k that are defined on marginal conditional probabilities of labels.**

The following points describe our main contributions.

Bayes (optimal) classifiers

We derive the form of the Bayes (optimal) classifier for the family of weighted instance-wise metrics that include precision@k or Hamming score@k, and their weighted variants, such as their propensity-scored versions. Following the literature, we include the results for recall@k [Menon et al., 2019] and (n)DCG [Jasinska and Dembczyński, 2018]. We do the same for label-wise metrics, including macro-averaged metrics under expected test utility (ETU) and population utility (PU) frameworks. Our findings demonstrate that across all considered metrics, the Bayes classifier can be defined based on marginal conditional probabilities of labels, allowing for the optimization of all these metrics using plug-in inference algorithms on top of existing label probability estimators. Considering that XMLC requires efficient algorithms, if necessary, we construct computationally efficient approximations that work in linear time with respect to the number of labels. For instance-wise metrics, an optimal decision rule boils down to simple reweighting of labels' probabilities and selection of top-k labels. However, for label-wise metrics, the inference procedure becomes more complex. Because of that, we propose an inference procedure based on the block coordinate ascend (BCA) algorithm for the ETU setting, and the algorithm based on the Frank Wolfe (FW) [Frank and Wolfe, 1956] method for finding the optimal classifier under the PU setting.

Regret bounds

We formally quantify the influence of the estimation error of the labels' marginal conditional probabilities on the suboptimality of the resulting classifier for all proposed inference algorithms. These results are expressed in the form of regret bounds [Bartlett et al., 2006, Narasimhan et al., 2015, Kotłowski and Dembczyński, 2017, Dembczyński et al., 2017]. The notion of consistency varies slightly across different statistical frameworks. A consistent classifier in the standard expected instance-wise utility (EIU) setting optimizes the expected utility on a single sample as the size of the training set increases. Under the ETU framework, the consistent classifier is the one that optimizes the expected prediction utility over a given test set. The PU framework focuses on estimation, in the sense that a consistent PU classifier is one that converges to the population optimal utility as the size of the training set increases. We show that for most metrics considered in this thesis, small estimation errors in label probabilities result in correspondingly small performance degradation, confirming the practical viability of our methods. Furthermore, under the assumption that the underlying method of estimating labels' marginal conditional probabilities is consistent (i.e., the estimation error decreases with increasing training data), our inference methods are also consistent, ensuring theoretical guarantees.

Efficient inference methods

The introduced inference methods have linear complexity in the number of labels, which is considered to be costly in the XMLC setting. Because of that, we demonstrate that by leveraging the sparsity of labels, we can reduce this complexity by orders of magnitude at a small cost of regret. We then examine probabilistic label trees (PLTs) [Jasinska et al., 2016], a popular family of XMLC classifiers that organize labels into a tree structure and factorize label probabilities over tree nodes. Although we also contribute consistency proofs for PLTs and propose their usage as the output layer of a neural network, our primary focus is to introduce PLT-based inference methods for finding exactly the top-k labels with reweighted probabilities characterized by sublinear complexity. We achieve this by applying the BF*(best-first-star) [Pearl, 1984] search algorithm as the PLT tree search procedure during inference.

Empirical evaluation

Finally, we confirm our theoretical findings with extensive empirical experiments on several popular XMLC benchmark datasets for the XMLC repository [Bhatia et al., 2016]. We first simulate the setting of perfect knowledge of the conditional marginal probabilities of labels, eliminating the main source of regret of all the methods and confirming their optimality We then conduct experiments using original labels to reflect the realistic scenario of imperfect probability estimates and the high data sparsity. Additionally, we demonstrate that the introduced methods can be used to optimize label-wise utilities that combine an instance-wise metric, favoring head labels, and a macro-averaged metric. This leads to a classifier that significantly improves tail-label performance with only a tiny sacrifice of head-label performance. Furthermore, this trade-off can be precisely controlled. Finally, we investigate the computational efficiency of a PLT inference method based on BF*-search, showing that it can be computationally less expensive compared to the simpler two-step strategy in which the prediction of marginal probabilities is followed by reweighting.

Related work

Optimization complex performance metrics

The problem of optimizing complex performance metrics is well-known, with many articles published for a variety of metrics and different classification problems. It has been considered for binary [Ye et al., 2012, Koyejo et al., 2014, Busa-Fekete et al., 2015, Dembczyński et al., 2017], multi-class [Narasimhan et al., 2015, 2022], multi-label [Waegeman et al., 2014, Koyejo et al., 2015, Kotłowski and Dembczyński, 2017], and multi-output [Wang et al., 2019b] classification.

Initially, the main focus was on designing algorithms, without a conscious emphasis on the statistical consequences of choosing models and their asymptotic behavior. Notable examples of such contributions are the SVMperf algorithm [Joachims, 2005], approaches suited to different types of the F-measure [Dembczyński et al., 2011, Natarajan et al., 2016, Jasinska et al., 2016], or precision at the top [Kar et al., 2015]. The wide use of such complex metrics has caused an increasing interest in investigating their theoretical properties, which can then serve as a guide to design practical algorithms.

The consistency of learning algorithms is a well-established problem. The seminal work of Bartlett et al. [2006] studied this problem for binary classification under the misclassification error. Since then a wide spectrum of learning problems and performance metrics has been analyzed in terms of consistency. These results concern ranking [Duchi et al., 2010, Ravikumar et al., 2011, Calauzenes et al., 2012, Yang and Koyejo, 2020], multi-class classification [Zhang, 2004, Tewari and Bartlett, 2007], multi-label classification [Koyejo et al., 2015, Kotłowski and Dembczyński, 2017] classification with abstention [Yuan and Wegkamp, 2010, Ramaswamy et al., 2018], or constrained classification problems [Agarwal et al., 2018, Kearns et al., 2018].

Our contribution is close to [Koyejo et al., 2015] and [Kotłowski and Dembczyński, 2017]. These articles analyze theoretically complex performance measures known from binary classification in the context of macro- and micro-averages in multi-label classification. However, they do not consider the predictions "at k".

Narasimhan et al. [2015] consider optimization of complex performance measures in multi-class setting and PU framework. In the follow-up article [Narasimhan et al., 2022], they extend the analysis to constraints defined by arbitrary functions of the confusion matrix. These constraints are, however, again different than the budgeted predictions in our case. We suspect that their approach can be generalized to our setting. And our Frank Wolfe algorithm was inspired by theirs.

On the other end of the spectrum are the budgeted at k instance-wise and micro-averaged metrics, such as precision@k or recall@k. The optimal classifiers for these metrics boil down to solving independent binary or multi-class problems via reduction like one-vs-all, pick-all-labels, pick-one-labels [Menon et al., 2019].

Efficient XMLC methods

In this thesis, we address not only the problem of making optimal predictions at k but also ensuring computational efficiency in the XMLC setting. Specifically, the inference methods we discuss can integrate efficiently with any label probability estimator capable of rapidly identifying a subset of top k', where $k' > k$, labels based on predicted marginal conditional probabilities of labels or scores that can be calibrated to probabilities. To provide context for our work, we present a brief overview of relevant methodologies in this domain. Please note that we provide here merely one simplified categorization among many possible perspectives on XMLC methods.

The past decade has witnessed the emergence of many diverse methods, which

aim to reduce computational costs and enhance scalability in XMLC problems. These include 1-vs-all-based approaches that leverage label sparsity [Yen et al., 2017, Babbar and Schölkopf, 2017], sophisticated label filtering methods [Vijayanarasimhan et al., 2014, Shrivastava and Li, 2015, Niculescu-Mizil and Abbasnejad, 2017], decision tree-based algorithms [Prabhu and Varma, 2014, Choromanska and Langford, 2015], and coding-based reduction strategies [Jasinska and Karampatziakis, 2016, Evron et al., 2018, Medini et al., 2019]. Recently, two families of methods have come to dominate the XMLC landscape: label tree methods and label embedding-based methods.

The label tree methods organize labels into a hierarchical tree structure, with individual labels placed at the leaves of this tree. Modern variants typically construct this tree via hierarchical clustering of labels, continuing until each cluster contains fewer than a few hundred labels. The tree structure over final label clusters is often called a router as additional classifiers in the inner nodes of the tree are tasked with guiding the inference into relevant clusters of labels. Classically, label tree approaches relied on sparse TF-IDF features [Leskovec et al., 2014] and trained separate linear classifiers at each tree node [Jasinska et al., 2016, Prabhu et al., 2018b, Khandagale et al., 2020, Jasinska-Kobus et al., 2020, Yu et al., 2022]. Over time, these models have evolved to incorporate more expressive architectures: starting with word and feature embeddings combined with a tree with linear models serving as both output and loss layer [Wydmuch et al., 2018], followed by LSTM-based models [Hochreiter and Schmidhuber, 1997] combined with an attention mechanism [Lin et al., 2017] in each tree node [You et al., 2019]. Most recently, encoder-only transformer architectures [Vaswani et al., 2017, Devlin et al., 2019] have been used to generate contextual representations of input instances [Chang et al., 2020, Ye et al., 2020, Zhang et al., 2021, Kharbanda et al., 2022a]. Notably, label tree methods operate under a standard classification framework and do not require any additional metadata or features for the labels beyond the training set of instance-labels samples. In this work, we discuss and extend probabilistic label trees [Jasinska et al., 2016], which are a special kind of label trees that estimate marginal conditional probabilities of labels by decomposing them on the paths from root to leaf in the tree.

A parallel line of research to label trees is focusing on label embedding-based methods. These approaches assume that the dimensionality of the label space has an implicit low-rank structureand seek to embed the feature space and label space into a joint low-dimensional space. Initial embedding-based approaches utilized simple linear projections [Mineiro and Karampatziakis, 2015, Yu et al., 2014, Bhatia et al., 2015]. Recognizing their limited expressiveness, subsequent approaches incorporated non-linear transformations [Tagami, 2017], and naturally started leveraging deep learning architectures [Yeh et al., 2017, Zhang et al., 2018, Wang et al., 2019a, Guo et al., 2019] utilizing autoencoders [Vincent et al., 2010] and dual encoders [Gillick et al., 2018]. For the inference, these deep learning-based methods rely on approximate nearest neighbor search algorithms [Malkov and Yashunin, 2018, Jayaram Subramanya et al., 2019]. The recent methods started taking advantage of the fact that labels often come with additional features like natural

language name or description, incorporating them through architectures like graph neural networks [Kipf and Welling, 2016] and encoder-only transformers [Zong and Sun, 2020, Saini et al., 2021]. To further enhance the predictive performance of embedding-based methods, they are often combined with additional per-label classifiers [Dahiya et al., 2021, Mittal et al., 2021, Saini et al., 2021, Dahiya et al., 2023a,b, Gupta et al., 2023].

The inference algorithms we introduce in this work are directly applicable to most of these XMLC methods. They work as plug-in approaches, relying on the estimates of the marginal conditional probabilities of labels. Most of the methods listed above either directly estimate these probabilities or predict a per-label score that can be calibrated. Furthermore, our approach can be combined with algorithmic approaches for dealing with missing labels [Saito et al., 2020, Qaraei et al., 2021] as well as with recently proposed methods that try to improve predictive performance on the tail labels by, e.g., leveraging label features to estimate label correlations [Mittal et al., 2021, Saini et al., 2021, Dahiya et al., 2021, Zhang et al., 2022] or to do data augmentation and generate new training points for tail labels [Wei et al., 2021, Kharbanda et al., 2022b, Chien et al., 2023].

Outline

This dissertation is organized into the following chapters:

- In Chapter 2, we formally introduce the setting of multi-label classification from the point of view of statistical decision theory. We then introduce the concept of performance metrics based on confusion matrices and introduce two frameworks under which they can be optimized: expected test utility (ETU) and Population Utility (PU). Then, we discuss the specific characteristics of XMLC, including long-tailed label distributions, prediction with the "at k" budget constraint, and the problem of missing labels.
- In Chapter 3, we analyze popular instance-wise metrics used in XMLC, including precision@k, recall@k, and DCG@k, as well as their weighted variants. For each metric, we derive the form of the optimal classifier and provide regret bounds in terms of the probability estimation errors.
- In Chapter 4, we extend the analysis of instance-wise metrics to the setting with missing (noisy) labels. We then critically examine the popular propensity model of Jain et al. [2016] (a model that estimates the probabilities that labels are missing). We also discuss the relationship between metrics using propensity models (so-called propensity-scored metrics) to estimate true performance on noisy data and tail label performance, motivating the need for developing and popularizing new metrics focusing on the performance of rare labels.
- In Chapter 5, we introduce label-wise utilities, and especially macro-averaged metrics, that belong to a family of performance metrics based on confusion

matrices and can serve as a good alternative for evaluating tail-label performance. We demonstrate that optimization of label-wise utilities under prediction constrained at k is a non-trivial optimization problem.

- In Chapter 6, we analyze the problem of optimization of label-wise metrics with budget @k under the ETU framework and show the form of the optimal classifier. We then propose an efficient approximate inference procedure based on the block coordinate ascent (BCA) algorithm and provide regret bounds.
- In Chapter 7, analogously to the previous chapter, we analyze the problem of optimization of label-wise metrics under the PU framework. We show the form of the optimal classifier, propose an algorithm for finding the optimal classifier using the Frank-Wolfe (FW) method, and provide regret bounds.
- In Chapter 8, we discuss efficient implementation strategies. First, we present a general method that leverages the sparsity of the label space to reduce the computational complexity of the introduced methods. Then, we introduce the probabilistic label tree (PLT), a popular classifier in XMLC, and show how to use its hierarchical structure for efficient inference for the discussed methods.
- In Chapter 9, we present empirical experiments that validate our theoretical findings and compare all algorithms introduced under the unified protocol.
- In Chapter 10, we conclude this dissertation with a brief summary.
- In Appendix A, we present the full derivation of the proofs that we found to be too long to present in the main body of this thesis.
- In Appendix B, we discuss additional technical details regarding empirical experiments and additional results.

Streszczenie

Ekstremalna klasyfikacja wieloetykietowa (ang. Extreme Multi-Label Classification, XMLC) to zadanie uczenia maszynowego (ang. machine learning) polegające na przypisywaniu do pojedynczego przykładu zbioru etykiet pochodzących z bardzo dużego zbioru, którego rozmiar może sięgać setek tysięcy, a nawet milionów możliwych etykiet. Z takimi problemami spotykamy się coraz częściej w wielu zastosowaniach uczenia maszynowego, takich jak tagowanie dokumentów i multimediów, systemy rekomendacyjne czy reklama internetowa. Jednak ekstremalna klasyfikacja wieloetykietowa napotyka poważne wyzwania, takie jak złożoność obliczeniowa czy problemy statystyczne wynikające z rzadkości występowania etykiet — zjawiska powszechnie określanego mianem problemu „długiego ogona" (ang. long-tail). Niniejsza rozprawa koncentruje się właśnie na etykietach pochodzących z „długiego ogona", które, mimo swojej rzadkości, mają duże znaczenie w wielu praktycznych zastosowaniach. Obecne miary jakości w ekstremalnej klasyfikacji wieloetykietowej, takie jak precyzja na k-tym miejscu (ang. precision@k), nie oddają jednak w wystarczającym stopniu jakości modeli uczenia maszynowego dla tych rzadkich etykiet, pozwalając klasyfikatorom je ignorować bez wpływu na wartość metryki. To motywuje potrzebę opracowania alternatywnego sposobu oceny skuteczności modeli względem etykiet z „długiego ogona". W związku z tym w niniejszej pracy zaproponowano wykorzystanie miar makro-uśrednianych (ang. macro-averaged metrics), które obliczają miary binarne osobno dla każdej etykiety, a następnie uśredniają wyniki, traktując wszystkie etykiety jako jednakowo ważne. Typowe zastosowania ekstremalnej klasyfikacji wieloetykietowej często wymagają przewidywania dokładnie k etykiet. To ograniczenie, w połączeniu z miarami makro-uśrednianymi, tworzy unikalne wyzwanie optymalizacyjne, które dotychczas nie było rozważane. Główna hipoteza badana w niniejszej rozprawie zakłada istnienie spójnych algorytmów wnioskowania umożliwiających optymalizację miar makro-uśrednianych przy ograniczeniu do k przewidywanych etykiet. Najważniejsze wkłady pracy obejmują wyprowadzenie klasyfikatorów Bayesowsko-optymalnych zarówno dla klasycznych metryk instancjowych, jak i dla miar makro-uśrednianych. Te ostatnie są analizowane w dwóch odmiennych

ramach statystycznych – oczekiwanej użyteczności testowej (ang. Expected Test Utility, ETU) oraz użyteczności populacyjnej (ang. Population Utility, PU). We wszystkich przypadkach przedstawiono ograniczenia żalu (ang. regret bounds), które gwarantują spójność w realistycznych warunkach, a także zaproponowano efektywne obliczeniowo algorytmy przewidywania etykiet. Wyniki empiryczne potwierdzają założenia teoretyczne, pokazując, że zaproponowane metody znacząco poprawiają wyniki w miarach makro-uśrednianych, jednocześnie utrzymując dobrą jakość w klasycznych metrykach i zachowując efektywność obliczeniową, co czyni je odpowiednimi do zastosowań w problemach ekstremalnej klasyfikacji wieloetykietowej.

Wprowadzenie i zarys

Uczenie maszynowe i ekstremalna klasyfikacja wieloetykietowa

Uczenie maszynowe (ang. machine learning) stanowi obecnie szeroką i dynamicznie rozwijającą się dziedzinę sztucznej inteligencji, znajdującą się na styku informatyki, optymalizacji matematycznej, teorii informacji, statystyki, a nawet neuronauki. Główną ideą uczenia maszynowego jest umożliwienie komputerom odkrywania algorytmów rozwiązujących zadania, które byłyby zbyt trudne lub kosztowne do rozwiązania przez ludzkich programistów. Zamiast otrzymywać szczegółowe instrukcje, komputer dostaje dane dotyczące problemu (np. przykłady poprawnie wykonanych zadań) i na ich podstawie „uczy się" rozwiązywać zadany problem.

Ostatnie postępy w rozwoju algorytmów uczenia maszynowego spowodowały, że znajduje ono zastosowanie w coraz większej liczbie aplikacji, co prowadzi do powstawania nowych kierunków badań. Podejścia uczenia maszynowego tradycyjnie dzieli się na trzy kategorie, odpowiadające różnym paradygmatom uczenia, w zależności od rodzaju danych dostępnych dla systemu:

- Uczenie nadzorowane: Komputer otrzymuje przykładowe dane wejściowe oraz oczekiwane wyniki (przykłady treningowe), dostarczone przez „nauczyciela". Zestaw takich przykładów nazywamy zbiorem treningowym. Celem jest nauczenie ogólnej funkcji mapującej dane wejściowe na wyjściowe.
- Uczenie nienadzorowane: Algorytm nie otrzymuje etykiet i samodzielnie identyfikuje ukryte struktury w danych wejściowych.
- Uczenie ze wzmocnieniem: Program komputerowy (zwykle zwany agentem) oddziałuje z dynamicznym środowiskiem, w którym musi realizować pewien cel. Podejmując różne akcje, otrzymuje informacje zwrotne analogiczne do nagród. Celem jest nauczenie strategii maksymalizującej sumę tych nagród.

Taki podział jest dość uproszczoną próbą skategoryzowania licznych podejść oraz algorytmów uczenia maszynowego. Niemniej jednak, w niniejszej pracy skupiamy się bezpośrednio na podklasie uczenia nadzorowanego, jaką jest klasyfikacja.

Klasyfikacja polega na przypisaniu jednej lub wielu dyskretnych etykiet do danej instancji (punktu w danych) na podstawie jej cech (np. identyfikacja obiektów widocznych na obrazie, diagnozowanie choroby na podstawie wyników pacjenta, przypisanie odpowiednich kategorii i tagów do artykułu, przewidywanie produktów, które mogą zostać zakupione przez konkretnego klienta). Klasyfikacja jest jednym z najszerzej badanych i stosowanych zadań uczenia maszynowego w licznych dziedzinach.

Proste zadania klasyfikacji zwykle obejmują przypisanie jednej etykiety (klasyfikacja wieloklasowa) lub określenie obecności bądź braku jednej cechy (klasyfikacja binarna). Jednak wiele rzeczywistych problemów wymaga jednoczesnego przypisania wielu etykiet jednej instancji. Taki rodzaj klasyfikacji, nazywany klasyfikacją wieloetykietową, wprowadza dodatkową złożoność. Na przykład artykuł prasowy może należeć jednocześnie do kategorii „polityka", „ekonomia" oraz „sprawy międzynarodowe"; obraz może zawierać jednocześnie wiele obiektów, jak „osoba", „samochód" czy „budynek"; klient może zakupić wiele produktów; pacjent może cierpieć na kilka schorzeń jednocześnie.

Ekstremalna klasyfikacja wieloetykietowa (ang. extreme multi-label classification, XMLC) rozszerza paradygmat klasyfikacji wieloetykietowej do przypadków z wyjątkowo dużą liczbą potencjalnych etykiet (od kilku tysięcy do wielu milionów), gdzie dla każdej instancji istotna jest tylko niewielka ich część. Taka sytuacja jest coraz bardziej powszechna w nowoczesnych zastosowaniach, takich jak:

1. Tagowanie dokumentów i mediów w wyszukiwarkach: Identyfikacja tematów, kategorii czy słów kluczowych z milionów możliwych tagów [Dekel and Shamir, 2010, Deng et al., 2011, Agrawal et al., 2013].
2. Systemach rekomendacyjnych: Proponowanie klientom niewielkiego podzbioru produktów z ogromnego katalogu na podstawie ich profili, przeglądanych przedmiotów [Weston et al., 2013, Zhuo et al., 2020, Song et al., 2020] lub zapytań do wyszukiwarki [Medini et al., 2019, Chang et al., 2020].
3. Reklama internetowa: Dopasowywanie reklam do zapytań lub stron internetowych z ich bardzo dużego zbioru. [Prabhu and Varma, 2014, Jain et al., 2019].

Skala tych problemów odróżnia klasyfikację ekstremalną od standardowej klasyfikacji wieloetykietowej, wymagając odmiennych technik i algorytmów. W XMLC zarówno przestrzeń cech, jak i etykiet mają wysoką wymiarowość, co rodzi kilka kluczowych problemów:

1. Wydajność obliczeniowa: metody klasyfikacji wieloetykietowej często charakteryzują się co najmniej liniową złożonością względem liczby etykiet lub większą, uwzględniając złożone zależności między etykietami. Zastosowanie tych metod staje się niepraktyczne przy dużej liczbie etykiet. W tej sytuacji niezbędne są wydajne algorytmy umożliwiające zarówno uczenie i inferencje

w czasie poniżej liniowego.

2. Problemy statystyczne: przy dużej liczbie potencjalnych etykiet większość z nich występuje jedynie w niewielkiej części instancji, prowadząc do poważnych problemów związanych z rzadkością i niezbalansowaniem danych. Duża pula etykiet sprawia również, że proces etykietowania jest podatny na błędy, zwłaszcza występuje tutaj problem brakujących etykiet, ponieważ jest praktycznie niemożliwe, aby osoby etykietujące dokładnie przejżała cały zbiór etykiet przy tak dużej ich liczbie.

Długi ogon etykiet

Chociaż w tej pracy dotykamy kwestii wydajności obliczeniowej algorytmów, szczególnie koncentrujemy się na problemie dużej nierównowagi i rzadkości etykiet. Biorąc pod uwagę dużą liczbę etykiet w zbiorach danych XMLC, sięgającą setek tysięcy czy nawet milionów, oraz ich zgodność z prawem Zipfa [Adamic and Huberman, 2002], nie jest zaskakujące, że wiele etykiet jest bardzo rzadkich, przez co rozkład etykiet charakteryzuje się tak zwanym „długim ogonem" [Bhatia et al., 2015, Babbar and Schölkopf, 2017]. Na figurze 1 przedstawiono rozkład częstości etykiet dla dziewięciu popularnych zbiorów danych z repozytorium XMLC [Bhatia et al., 2016], co wyraźnie pokazuje, że tylko niewielka część etykiet jest dobrze reprezentowana w całym zbiorze, podczas gdy pozostałe tworzą długi ogon, mając niewiele przykładów.

Paradoksalnie, w wielu zastosowaniach te rzadkie etykiety są uważane za ważniejsze, jako bardziej informacyjne lub satysfakcjonujące w przypadku poprawnej predykcji; przykładowo, bardzo ogólne tagi są rzadko użyteczne podczas wyszukiwania informacji, a użytkownik czuje większą satysfakcję, gdy otrzymuje trafną rekomendację nieznanego wcześniej przedmiotu zamiast popularnego, o którym już wie. Powstaje więc istotne pytanie badawcze: jak najlepiej uwzględnić te rzadkie etykiety?

Miary jakości w XMLC

Podczas uczenia klasyfikatora zwykle dąży się do uzyskania jak najwyższej jakości predykcji, wyrażonej w formie metryki oceniającej wyjście klasyfikatora względem oczekiwanego wyjścia. Do oceny zdolności klasyfikatora do generalizacji używa się dodatkowego zbioru (testowego), który nie był wcześniej widziany podczas uczenia klasyfikatora. Zbiór testowy, podobnie jak treningowy, musi zawierać zarówno dane wejściowe, jak i odpowiadające im poprawne etykiety. Najprostszą metryką w klasyfikacji jest dokładność zero-jedynkowa (ang. zero-one accuracy), przypisująca wartość 1, jeśli predykcja klasyfikatora dokładnie odpowiada oczekiwanemu wyjściu

dla danego wejścia, a 0 w przeciwnym przypadku. Wartości te są następnie uśredniane po wszystkich instancjach w zbiorze testowym.

Ze względu na specyfikę zagadnienia, obszar XMLC przyjął własny zestaw standardowych miar do oceny klasyfikatorów. W klasycznej klasyfikacji wieloetykietowej ocena decyzji klasyfikatora odbywa się dla każdej etykiety osobno lub dla całego zbioru etykiet pozytywnych i negatywnych. Z powodu charakterystyki aplikacji XMLC, takich jak wyszukiwanie i rekomendacje, często nie jest istotne uzyskanie dokładnej decyzji dla każdej etykiety, lecz raczej zestawu etykiet określonej wielkości, który maksymalizuje daną metrykę oceny. Wielkość zestawu predykcji często wynika z interfejsu graficznego użytkownika, mającego określoną liczbę miejsc na wyniki wyszukiwania, rekomendacje lub reklamy [Cremonesi et al., 2010, Chang et al., 2021]. Z tego powodu popularne są metryki typu na „k-tym miejscu" (ang. „at k", @k), gdzie klasyfikator zwraca dokładnie k etykiet (czasami również z określoną kolejnością). Najpopularniejszą metryką jest precyzja@k (ang. precision@k), licząca, ile spośród k przewidzianych etykiet należy do zbioru poprawnych etykiet, a następnie metryki takie jak czułość@k (ang. recall@k) czy (n)DCG@k.

Zauważono, że przy dużej nierównowadze częstości występowania etykiet, dobre wyniki można osiągnąć, przewidując najczęściej występujące etykiety [Jain et al., 2016, Wei and Li, 2020]. W efekcie społeczność zaczęła rozwijać nowe metryki premiujące „wartościowe"[Ye et al., 2020], „różnorodne"[Babbar and Schölkopf, 2019] i „rzadkie oraz informacyjne" [Prabhu et al., 2018a] etykiety kosztem najpopularniejszych.

Problem ten jako pierwsi podjęli Jain et al. [2016], wprowadzając miary oparte o skłonność etykiet do bycia zaobserwowanymi (ang. propensity scores), które szybko zyskały popularność i stały się standardem w społeczności XMLC. Są to warianty ważone metryk standardowych, jak precyzja@k, premiujące rzadkie etykiety, jednak interpretacja wyników tych metryk nie jest oczywista.

Motywacja

Zauważyliśmy, że nawet popularna precyzja ważona skłonnością (ang. propensity-scored precision) jedynie nieznacznie lepiej różnicuje klasyfikatory, które skutecznie przewidują rzadkie etykiety, od tych,które sobie z nimi nie radzą. Ta obserwacja motywuje poszukiwanie alternatywnych miar jakości, lepiej nadających się do oceny jakości klasyfikatora na etykietach z długiego ogona.

W klasyfikacji wieloetykietowej można rozważać szerokie spektrum miar, które zwykle dzieli się na trzy kategorie, zależne od sposobu uśredniania wyników: uśredniane po instancjach (ang. instance-wise), mikro- (ang. micro-averaging) oraz makro-uśrednianych (ang. macro-averaging). Miary uśredniane po instancjach są zdefiniowane, jak sugeruje nazwa, na poziomie pojedynczej instancji. Typowymi przykładami są wcześniej wspomniane precyzja@k, czułość@k, (n)DCG@k, strata Hamminga oraz instancyjna miara F_β. Mikro-uśrednianie definiuje się na poje-

dynczej macierzy pomyłek, która sumuje prawdziwie pozytywne, fałszywie pozytywne, fałszywie negatywne oraz prawdziwie negatywne wyniki ze wszystkich etykiet. Makro-uśrednianie wymaga, aby binarna miara została zastosowana osobno dla każdej etykiety, a następnie uśredniona po wszystkich etykietach. Generalnie każdą binarną miarę można zastosować w każdym ze wspomnianych schematów uśredniania. Nic więc dziwnego, że niektóre miary, np. strata Hamming, prowadzą do tej samej postaci końcowej niezależnie od użytego schematu. Można również rozważać szerszą klasę miar, definiowanych jako ogólne funkcje agregujące macierze pomyłek poszczególnych etykiet.

Metryki makro-uśrednione wydają się szczególnie atrakcyjne w kontekście oceny skuteczności w problemach XMLC o długim ogonie, ponieważ traktują wszystkie etykiety jako równie ważne, zapobiegając marginalizacji tych, które posiadają niewiele przykładów pozytywnych. Z tego powodu dokładniej analizujemy te metryki przy budżecie predykcji k, który naturalnie pasuje do ustawienia XMLC. Budżet ten wymaga od algorytmu predykcyjnego „rozsądnego" wyboru etykiet, co często powoduje, że wśród przewidywanych etykiet znajdują się również te należące do długiego ogona. Jest to również naturalne w typowych zastosowaniach XMLC, takich jak systemy rekomendacyjne, w których liczba pozycji jest zdeterminowana interfejsem użytkownika.

Aby zilustrować nacisk, jaki metryki makro-uśrednione kładą na rzadkie etykiety, w tabeli 1 porównujemy dwa klasyfikatory:[3] Pierwszy był trenowany na wszystkich etykietach, a drugi jedynie na 20% najczęstszych etykiet (pozostałe odrzucono). Oba modele oceniano na pełnym zbiorze etykiet. Standardowe metryki wykazują jedynie niewielką zmianę po redukcji przestrzeni etykiet do najczęstszych. Precyzja ważona skłonnością spada zauważalnie bardziej, czasem nawet do 35%, jednak ten spadek nie oddaje w pełni niezdolności klasyfikatora do przewidywania pozostałych 80% etykiet. W przeciwieństwie do tego makro-uśrednione metryki (precyzja, czułość oraz miara F_1) przy budżecie k, a także pokrycie@k (ang. coverage@k), popularna miara pomocnicza w XMLC [Jain et al., 2016, Babbar and Schölkopf, 2019, Wei and Li, 2020, Schultheis et al., 2022], spadają znacznie silniej, nawet ponad 70%, znacznie lepiej oddając brak zdolności klasyfikatora do przewidywania pozostałych 80% etykiet. Wyniki te wskazują, że makro-miary z budżetem k mogą być bardzo atrakcyjne w kontekście długiego ogona.

W niniejszej pracy analizujemy problem rzadkich etykiet przez pryzmat miar jakości do ewaluacji klasyfikatorów. Proponujemy użycie makro-uśrednionych metryk z budżetem predykcji k jako alternatywy, koncentrując się na skuteczności przewidywania „rzadkich etykiet", łatwych w interpretacji, i badamy problem ich optymalizacji. Interesującym jest fakt, że optymalizację makro-uśrednionych metryk można rozważać w dwóch odrębnych podejściach statystycznych – oczekiwanej użyteczność na zbiorze testowym (ang. expected test utility, ETU) oraz użyteczności populacyjnej (ang. population utility, PU) [Ye et al., 2012, Dembczyński et al., 2017]. Podejście ETU polega na optymalizacji wartości oczekiwanej

[3] Używamy zespołu probabilistycznych drzew etykiet [Jasinska-Kobus et al., 2020], tej samej metody, którą później stosujemy w rodziale 9. do reszty eksperymentów.

metryki na danym zbiorze testowym, podczas gdy PU dąży do optymalizacji wartości metryki obliczonej na oczekiwanej macierzy pomyłek całej populacji. O ile makro-uśrednione miary, a także ogólniejsza klasa miar oparta na macierzach pomyłek poszczególnych etykiet, były już dobrze zbadane w klasyfikacji wieloetykietowej w obu podejściach, o tyle warianty z budżetem k nie były dotychczas analizowane. Ponadto wcześniej wprowadzone podejścia optymalizacyjne nie mogą być bezpośrednio zastosowane do problemów z budżetem, ponieważ ograniczenie do k predykcji, silnie wiąże ze sobą etykiety, czyniąc problem optymalizacji znacznie trudniejszym.

Jednym z pierwszych wyników tej pracy jest obserwacja, że istnieje klasa metryk liczonych po etykietach z budżetem k, które można liniowo rozłożyć na poszczególne etykiety w formie binarnych użyteczności. Klasa ta obejmuje zarówno standardowe metryki uśredniane po instanacjach, jak precision@k, jak i metryki makro-uśrednione, potencjalnie umożliwiając uzyskanie ogólnych wyników w ramach jednolitego podejścia. Z tego powodu interesuje nas zbadanie, czy istnieją spójne algorytmy inferencji optymalizujące metryki liczone po etykietach z budżetem k oraz czy są one kompatybilne z istniejącymi podejściami do XMLC. Ze względu na bardzo dużą przestrzeń etykiet, niemal wszystkie istniejące metody XMLC dążą do wygenerowania najlepszych k kandydatów wraz z przypisanymi do poszczególnych etykiet ocenami ich istotności. Najlepiej, aby oceny te odzwierciedlały (lub mogły być skalibrowane do) estymaty brzegowych prawdopodobieństw warunkowych etykiet. Tak więc, aby algorytm optymalizujący metryki liczonych po etykietach ograniczone do k był praktyczny, powinien operować na brzegowych prawdopodobieństwach warunkowych etykiet. Pozwoliłoby to na zastosowanie go do istniejących i przyszłych metod, niezależnie od ich architektury. Dodatkowo, oczekujemy od takiego algorytmu spójności (ang. consistency), co oznacza, że wraz ze wzrostem rozmiaru zbioru treningowego, żal (ang. regret) klasyfikatora maleje do zera.

Cel i zakres pracy

Biorąc pod uwagę przedstawione powyżej motywacje, hipoteza niniejszej rozprawy została sformułowana następująco:

> **Istnieją spójne algorytmy inferencyjne do optymalizacji miar liczonych po etykietach z ograniczeniem predykcji do k-tego miejsca, które można zdefiniować za pomocą marginalnych warunkowych prawdopodobieństw etykiet.**

Poniżej przedstawiamy główne osiągnięcia tej pracy.

Klasyfikatory Bayesowskie (optymalne)

Wyprowadzamy postać klasyfikatora Bayesowskiego (optymalnego) dla rodziny ważonych miar uśrednianych po instancjach, obejmujących precyzje@k lub stratę

Hamminga@k, a także ich wersji ważonych, takich jak ich odmiany ważone skłonnością. Idąc za literaturą, uwzględniamy wyniki dla czułości@k [Menon et al., 2019] oraz (n)DCG [Jasinska and Dembczyński, 2018]. Podobne wyniki przedstawiamy dla miar dla etykiet, w tym metryk uśrednianych makro, w ramach podejść opartych o oczekiwaną użyteczność na zbiorze testowym (ETU) oraz użyteczność populacyjną (PU). Nasze wyniki pokazują, że dla wszystkich rozważanych miar klasyfikator Bayesowski można wyrazić przez marginalne warunkowe prawdopodobieństwa etykiet, co umożliwia optymalizację tych metryk wykorzystując istniejące podejścia do estymowania prawdopodobieństw etykiet. Ponieważ klasyfikacja XMLC wymaga wydajnych algorytmów, w razie potrzeby konstruujemy aproksymacje obliczeniowe działające w czasie liniowym względem liczby etykiet. Dla metryk uśrednianych po instancjach optymalna reguła decyzyjna sprowadza się do prostego przeważenia prawdopodobieństw etykiet i wybrania k najwyższych. Jednak dla metryk liczonych po etykietach procedura inferencyjna jest bardziej złożona. W związku z tym proponujemy procedurę inferencji opartą na algorytmie optymalizacji według współrzędnych blokowych (ang. block coordinate ascent, BCA) dla podejścia ETU oraz algorytm oparty na metodzie Franka-Wolfe'a (FW) [Frank and Wolfe, 1956] w celu znalezienia optymalnego klasyfikatora w ramach podejścia PU.

Ograniczenia na żal klasyfikatorów

Formalnie kwantyfikujemy wpływ błędu estymacji marginalnych warunkowych prawdopodobieństw etykiet na suboptymalność wynikowego klasyfikatora dla wszystkich proponowanych algorytmów inferencyjnych. Wyniki te wyrażamy w postaci ograniczeń żalu klasyfikatora [Bartlett et al., 2006, Narasimhan et al., 2015, Kotłowski and Dembczyński, 2017, Dembczyński et al., 2017]. Pojęcie spójności różni się nieznacznie w zależności od rozważanych ram statystycznych. Zgodny klasyfikator w standardowym ujęciu oczekiwanej użyteczności na instancji (EIU, ang. expected instance utility) optymalizuje oczekiwaną użyteczność na próbce wraz ze wzrostem rozmiaru zbioru treningowego. W podejściu ETU zgodnym klasyfikatorem jest ten, który optymalizuje oczekiwaną użyteczność predykcji na ustalonym zbiorze testowym. Natomiast podejście PU koncentruje się na estymacji, w tym sensie, że zgodny klasyfikator PU to taki, który wraz ze wzrostem rozmiaru zbioru treningowego zbiega do optymalnej użyteczności populacyjnej. Pokazujemy, że dla większości miar rozważanych w tej pracy, niewielkie błędy w estymacji prawdopodobieństw etykiet prowadzą do odpowiednio niewielkiego pogorszenia wyników, co potwierdza praktyczną użyteczność proponowanych metod. Ponadto, przy założeniu, że użyta metoda estymacji marginalnych warunkowych prawdopodobieństw etykiet jest spójna (tj. błąd estymacji maleje wraz ze wzrostem danych treningowych), proponowane metody inferencyjne również są zgodne, zapewniając tym samym gwarancje teoretyczne.

Wydajne metody inferencyjne

Przedstawione metody inferencyjne mają złożoność liniową względem liczby etykiet, co w kontekście XMLC jest uznawane za kosztowne. Dlatego pokazujemy, że wyko-

rzystując rzadkość etykiet, można ograniczyć tę złożoność o rzędy wielkości kosztem niewielkiego zwiększenia żalu klasyfikatora. Badamy następnie probabilistyczne drzewa etykiet (ang. probabilistic label trees, PLTs) [Jasinska et al., 2016], popularną rodzinę klasyfikatorów XMLC, które organizują etykiety w strukturę drzewa i faktoryzują prawdopodobieństwa etykiet na węzłach drzewa. Mimo że dostarczamy również dowodów spójności dla PLT oraz proponujemy ich wykorzystanie jako warstwę wyjściową sieci neuronowej, naszym głównym celem jest przedstawienie metod inferencyjnych opartych na PLT o złożoności podliniowej, do znajdowania dokładnie k etykiet z największymi prawdopodbieństw wymnośonych razy dostarczone wagi. Osiągamy to, stosując algorytm BF* (ang. best-first-star) [Pearl, 1984] jako procedurę przeszukiwania drzewa PLT podczas inferencji.

Ewaluacja empiryczna

Na koniec, potwierdzamy nasze wyniki teoretyczne szeroko zakrojonymi eksperymentami empirycznymi na kilku popularnych zbiorach danych XMLC z repozytorium XMLC [Bhatia et al., 2016]. Najpierw symulujemy scenariusz perfekcyjnej znajomości marginalnych warunkowych prawdopodobieństw etykiet, eliminując główne źródło żalu wprowadzonych metod i potwierdzając ich optymalność. Następnie przeprowadzamy eksperymenty z oryginalnymi etykietami, odzwierciedlając realistyczny scenariusz niedoskonałej estymacji prawdopodobieństw oraz dużej rzadkości danych. Dodatkowo pokazujemy, że wprowadzone metody umożliwiają optymalizację użyteczności dla etykiet, które łączą metrykę instancyjną, preferującą popularne etykiety, z metryką makro-uśrednianą. Pozwala to znacząco poprawić wyniki na rzadkich etykietach kosztem niewielkiego spadku jakości na popularnych etykietach, a kompromis ten można precyzyjnie kontrolować. Na koniec badamy wydajność obliczeniową metody inferencyjnej PLT opartej na wyszukiwaniu BF*, pokazując, że może być ona obliczeniowo tańsza w porównaniu do prostszej strategii dwuetapowej.

Powiązane prace

Optymalizacja złożonych metryk wydajności

Problem optymalizacji złożonych metryk wydajności jest dobrze znany, istnieje duża liczba publikacji analizująca szeroki zakres miar dla różnych problemów klasyfikacji. Problem ten rozważany był dla klasyfikacji binarnej [Ye et al., 2012, Koyejo et al., 2014, Busa-Fekete et al., 2015, Dembczyński et al., 2017], wieloklasowej [Narasimhan et al., 2015, 2022], wieloetykietowej [Waegeman et al., 2014, Koyejo et al., 2015, Kotłowski and Dembczyński, 2017], oraz wielo-wyjściowej [Wang et al., 2019b].

Początkowo główny nacisk kładziono na projektowanie algorytmów, bez świadomego uwzględnienia konsekwencji statystycznych wyboru modeli i ich zachowania asymptotycznego. Przykładami takich podejść są algorytm SVMperf [Joachims, 2005],

metody dedykowane różnym typom miary F [Dembczyński et al., 2011, Natarajan et al., 2016, Jasinska et al., 2016] lub precyzji w czołówce rankingu [Kar et al., 2015]. Rosnące zastosowanie tak złożonych metryk spowodowało zainteresowanie ich teoretycznymi własnościami, które mogą służyć jako wskazówki przy projektowaniu praktycznych algorytmów.

Spójność algorytmów uczenia jest dobrze zbadanym zagadnieniem. Przełomowa praca Bartlett et al. [2006] badała ten problem dla klasyfikacji binarnej pod względem błędu klasyfikacji. Od tego czasu przeanalizowano już szerokie spektrum problemów uczenia oraz miar wydajności pod kątem spójności. Wyniki te obejmują rankingi [Duchi et al., 2010, Ravikumar et al., 2011, Calauzenes et al., 2012, Yang and Koyejo, 2020], klasyfikację wieloklasową [Zhang, 2004, Tewari and Bartlett, 2007], wieloetykietową [Koyejo et al., 2015, Kotłowski and Dembczyński, 2017], klasyfikację z możliwością wstrzymania się od decyzji [Yuan and Wegkamp, 2010, Ramaswamy et al., 2018] oraz problemy klasyfikacji z ograniczeniami [Agarwal et al., 2018, Kearns et al., 2018].

Wkład tej pracy jest bliski [Koyejo et al., 2015] oraz [Kotłowski and Dembczyński, 2017]. Artykuły te analizują teoretycznie złożone miary wydajności znane z klasyfikacji binarnej w kontekście makro- i mikro-uśredniania w klasyfikacji wieloetykietowej. Nie rozważają one jednak predykcji „na k-tym miejscu". Narasimhan et al. [2015] rozważają optymalizację złożonych miar wydajności w klasyfikacji wieloklasowej oraz w ramach użyteczności populacyjnej (PU).

W kolejnym artykule [Narasimhan et al., 2022] rozszerzają analizę na ograniczenia zdefiniowane przez dowolne funkcje macierzy pomyłek. Te ograniczenia są jednak inne niż budżetowe predykcje rozważane przez nas. Przypuszczamy, że ich podejście może być uogólnione na nasz przypadek. Nasz algorytm Franka-Wolfe'a został zainspirowany ich podejściem.

Na drugim końcu spektrum znajdują się metryki uśredniane po instancjach i mikro-uśredniane typu „@k", takie jak precyzja@k lub czułość@k. Optymalne klasyfikatory dla tych metryk sprowadzają się do rozwiązywania niezależnych problemów binarnych lub wieloklasowych poprzez redukcje typu jeden przeciwko wszystkim (ang. one-vs-all), wybór wszystkich etykiet (ang. pick-all-labels), czy wybór jednej etykiety (ang. pick-one-labels) [Menon et al., 2019].

Efektywne metody XMLC

W niniejszej pracy podejmujemy nie tylko problem optymalnego dokonywania predykcji dla k etykiet, lecz także zapewnienia efektywności obliczeniowej w kontekście ekstremalnej klasyfikacji wieloetykietowej (XMLC). Metody inferencji, które omawiamy, mogą być efektywnie integrowane z dowolnym estymatorem prawdopodobieństw etykiet, zdolnym do szybkiego wskazania podzbioru k' najlepszych etykiet na podstawie przewidywanych brzegowych prawdopodobieństw warunkowych etykiet lub ocen, które mogą być skalibrowane do postaci prawdopodobieństw. Aby zapewnić kontekst dla naszych rozważań, przedstawiamy krótki przegląd istotnych metodologii w tej dziedzinie. Należy zauważyć, że przedstawiona tutaj kategoryzacja jest tylko jedną spośród wielu możliwych perspektyw

na metody XMLC.

Ostatnia dekada zaowocowała wieloma różnorodnymi metodami mającymi na celu redukcję kosztów obliczeniowych oraz poprawę skalowalności problemów XMLC. Należą do nich metody oparte na podejściu jeden-przeciwko-wszystkim (1-vs-all), które wykorzystują rzadkość etykiet [Yen et al., 2017, Babbar and Schölkopf, 2017], zaawansowane metody filtrowania etykiet [Vijayanarasimhan et al., 2014, Shrivastava and Li, 2015, Niculescu-Mizil and Abbasnejad, 2017], algorytmy oparte na drzewach decyzyjnych [Prabhu and Varma, 2014, Choromanska and Langford, 2015] oraz strategie redukcji oparte na kodowaniu [Jasinska and Karampatziakis, 2016, Evron et al., 2018, Medini et al., 2019]. Ostatnio dwie rodziny metod zaczęły dominować krajobraz XMLC: metody oparte na drzewa etykiet (ang. label tree) oraz metody oparte na zanurzeniach etykiet (ang. label embedding).

Metody operate o drzewa etykiet organizują etykiety w hierarchiczną strukturę drzewa, z pojedynczymi etykietami umieszczonymi w liściach drzewa. Nowoczesne warianty zazwyczaj konstruują to drzewo poprzez hierarchiczną klasteryzację etykiet, kontynuowaną do momentu, gdy każdy klaster zawiera mniej niż kilka-set etykiet. Struktura drzewa obejmująca klastry etykiet nazywana jest często routerem, ponieważ dodatkowe klasyfikatory w węzłach wewnętrznych drzewa mają za zadanie prowadzić proces inferencji do właściwych klastrów etykiet. Klasyczne podejścia drzew etykiet wykorzystywały rzadkie cechy TF-IDF [Leskovec et al., 2014] i trenowały oddzielne liniowe klasyfikatory w każdym węźle drzewa [Jasinska et al., 2016, Prabhu et al., 2018b, Khandagale et al., 2020, Jasinska-Kobus et al., 2020, Yu et al., 2022]. Z czasem modele te ewoluowały i zaczęły korzystać z bardziej złożonych architektur: początkowo wykorzystywano proste zanurzenia słów lub cech w połączeniu z drzewem zawierającym modele liniowe, które służyło jako warstwa wyjściowa jak i funkcja straty [Wydmuch et al., 2018], następnie pojawiły się modele oparte na rekurencyjnych sieciach neuronowych (ang. recurrent neural networks) [Hochreiter and Schmidhuber, 1997] wzbogacone o mechanizm uwagi (ang. attention) [Lin et al., 2017] w każdym węźle drzewa [You et al., 2019]. Najnowsze podejścia wykorzystują enkodery oparte na transformerach (ang. encoder-only transformers) [Vaswani et al., 2017, Devlin et al., 2019] w celu generowania kontekstowych reprezentacji instancji wejściowych [Chang et al., 2020, Ye et al., 2020, Zhang et al., 2021, Kharbanda et al., 2022a]. Co istotne, metody drzew etykiet działają w standardowym paradygmacie klasyfikacji i nie wymagają dodatkowych metadanych ani cech etykiet poza zbiorem treningowym zawierającym pary instancja-etykieta. W tej pracy omawiamy i rozszerzamy probabilistyczne drzewa etykiet [Jasinska et al., 2016], stanowiące szczególny rodzaj drzew etykiet, które szacują brzegowe prawdopodobieństwa warunkowe etykiet poprzez ich dekompozycję na ścieżkach od korzenia do liści drzewa.

Równoległą linią badań względem drzew etykiet są metody oparte na zanurzeniach etykiet. Podejścia te zakładają, że przestrzeń etykiet posiada ukrytą niskowymiarową strukturę i dążą do zanurzenia przestrzeni cech i etykiet we wspólnej przestrzeni niskowymiarowej. Początkowe podejścia do osadzania stosowały proste projekcje liniowe [Mineiro and Karampatziakis, 2015, Yu et al., 2014, Bhatia et al., 2015]. Rozpoznając ich ograniczoną ekspresyjność, kolejne podejścia

wprowadziły transformacje nieliniowe [Tagami, 2017], oraz zaczęły naturalnie wykorzystywać architektury głębokiego uczenia [Yeh et al., 2017, Zhang et al., 2018, Wang et al., 2019a, Guo et al., 2019], takie jak autoenkodery [Vincent et al., 2010] i podwójne enkodery [Gillick et al., 2018]. Metody te korzystają z przybliżonych algorytmów najbliższych sąsiadów (ang. nearest neighbor) do inferencji [Malkov and Yashunin, 2018, Jayaram Subramanya et al., 2019]. Najnowsze podejścia zaczęły wykorzystywać dodatkowe cechy etykiet, takie jak nazwy czy opisy w języku natrulanym, integrując je poprzez grafowe sieci neuronowe [Kipf and Welling, 2016] oraz transformery [Zong and Sun, 2020, Saini et al., 2021]. Dla dalszej poprawy skuteczności metody te często są łączone z dodatkowymi klasyfikatorami na poziomie etykiet [Dahiya et al., 2021, Mittal et al., 2021, Saini et al., 2021, Dahiya et al., 2023a,b, Gupta et al., 2023].

Metody inferencji, które wprowadzamy, są bezpośrednio aplikowalne do większości wymienionych metod XMLC, opierając się na oszacowaniach marginalnych prawdopodobieństw warunkowych etykiet, które pochodzą od klasyfikatora. Większość powyższych metod albo bezpośrednio estymuje te prawdopodobieństwa, albo przewiduje wyniki dla poszczególnych etykiet, które można skalibrować. Ponadto nasze podejście można łączyć z algorytmami radzącymi sobie z brakującymi etykietami [Saito et al., 2020, Qaraei et al., 2021] oraz z metodami poprawiającymi wydajność predykcyjną dla rzadkich etykiet poprzez np. wykorzystanie cech etykiet do oszacowania ich korelacji [Mittal et al., 2021, Saini et al., 2021, Dahiya et al., 2021, Zhang et al., 2022] lub poprzez augmentację danych i generowanie nowych przykładów treningowych dla rzadkich etykiet [Wei et al., 2021, Kharbanda et al., 2022b, Chien et al., 2023].

Zarys pracy

Niniejsza rozprawa składa się z następujących rozdziałów:

- W rozdziale 1., który został przetłumaczony powyżej, zerysowujemy problematykę uczenia maszynowego, klasyfikacji wieloetykietowej i klasyfikacji ekstramalnej. Omawiamy problem długiego ogona etykie i mierzenia trafnośći przewidywania przez klasyfikatory tych rzadkich etykiet, który stanowi główne zagadnienie tej pracy. Krótko omawiamy wkład tej rozprawy, oraz powiązaną literaturę.
- W rozdziale 2. formalnie wprowadzamy problem klasyfikacji wieloetykietowej z punktu widzenia statystycznej teorii decyzji. Następnie omawiamy metryki jakości oparte o macierz pomyłek oraz przedstawiamy dwie ramy ich optymalizacji: oczekiwaną użyteczność testową (ETU) oraz użyteczność populacyjną (PU). Następnie omawiamy szczególne cechy klasyfikacji ekstremalnie wieloetykietowej (XMLC), takie jak rozkłady etykiet o długim ogonie, predykcja przy ograniczeniu „@k” oraz problem brakujących etykiet.
- W rozdziale 3. analizujemy popularne miary stosowane w XMLC, uśredni-

ane po instancja, w tym precyzje@k, czułość@k oraz DCG@k, jak również ich ważone warianty. Dla każdej z metryk wyprowadzamy postać optymalnego klasyfikatora oraz podajemy ograniczenia żalu wyrażone w kategoriach błędów estymacji prawdopodobieństw.

- W rozdziale 4. rozszerzamy analizę metryk uśrednianych po instancjach do setting brakujących (zaszumionych) etykiet. Następnie krytycznie analizujemy popularny model skłonności (ang. propensity model) zaproponowany przez Jain et al. [2016], służący do estymowania prawdopodobieństw brakujących etykiet. Omawiamy również związek między metrykami stosującymi modele skłonności (tzw. metrykami z poprawką skłonnościową) do szacowania rzeczywistej jakości klasyfikatora na zaszumionych danych a skutecznością klasyfikacji dla etykiet z długiego ogona, motywując potrzebę opracowania i popularyzacji nowych metryk skupionych na ocenie jakości klasyfikacji rzadkich etykiet.
- W rozdziale 5. wprowadzamy użyteczności liczone po etykietach, zwłaszcza metryki uśrednione makro, należące do rodziny metryk jakości opartych na macierzach pomyłek, które mogą stanowić dobrą alternatywę do oceny jakości klasyfikacji rzadkich etykiet. Pokazujemy, że optymalizacja metryk etykietowych przy ograniczeniu predykcji do k jest nietrywialnym problemem optymalizacyjnym.
- W rozdziale 6. analizujemy problem optymalizacji metryk etykietowych z ograniczeniem liczby predykcji „@k" w ramach ETU oraz wyprowadzamy postać optymalnego klasyfikatora. Następnie proponujemy efektywną procedurę aproksymacji wnioskowania opartą na algorytmie optymalizacji według współrzędnych blokowych (BCA) oraz przedstawiamy oszacowania żalu.
- W rozdziale 7., analogicznie do poprzedniego rozdziału, analizujemy problem optymalizacji metryk etykietowych w ramach PU. Pokazujemy postać optymalnego klasyfikatora, proponujemy algorytm do znajdowania optymalnego klasyfikatora metodą Franka-Wolfe'a (FW) oraz podajemy oszacowania żalu.
- W rozdziale 8. omawiamy efektywne strategie implementacyjne. Najpierw przedstawiamy ogólną metodę wykorzystującą rzadkość przestrzeni etykiet do redukcji złożoności obliczeniowej wprowadzonych metod. Następnie wprowadzamy probabilistyczne drzewa etykiet (PLT), popularny klasyfikator XMLC, oraz pokazujemy jak wykorzystać ich hierarchiczną strukturę do efektywnego wnioskowania dla omawianych metod.
- W rozdziale 9. przedstawiamy eksperymenty empiryczne, które weryfikują nasze wyniki teoretyczne oraz porównują wszystkie wprowadzone algorytmy według jednolitego protokołu.
- W rozdziale 10. kończymy niniejszą rozprawę krótkim podsumowaniem.
- W załączniku A zamieszczamy pełne wyprowadzenia dowodów, które były zbyt długie, by umieścić je w głównej części pracy.
- W załączniku B omawiamy dodatkowe szczegóły techniczne dotyczące przeprowadzonych eksperymentów oraz przedstawiamy dodatkowe wyniki.

Bibliography

L. A. Adamic and B. A. Huberman. Zipf's law and the internet. Glottometrics, 3 (1):143–150, 2002.

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In J. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 60–69. PMLR, 10–15 Jul 2018.

R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pages 13–24. International World Wide Web Conferences Steering Committee / ACM, 2013.

R. Babbar and B. Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. In Proceedings of the tenth ACM international conference on web search and data mining, pages 721–729, New York, NY, USA, 2017. Association for Computing Machinery.

R. Babbar and B. Schölkopf. Data scarcity, robustness and extreme multi-label classification. Machine Learning, 108(8), 09 2019. doi: 10.1007/s10994-019-05791-5.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473):138–156, 2006.

K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain. Sparse local embeddings for extreme multi-label classification. In Advances in Neural Information Processing Systems 28, pages 730–738. Curran Associates, Inc., 2015.

K. Bhatia, K. Dahiya, H. Jain, A. Mittal, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets and code, 2016. URL `http://manikvarma.org/downloads/XC/XMLRepository.html`.

R. Busa-Fekete, B. Szörényi, K. Dembczyński, and E. Hüllermeier. Online f-measure optimization. In Advances in Neural Information Processing Systems 28, pages 595–603. Curran Associates, Inc., 2015.

C. Calauzenes, N. Usunier, and P. Gallinari. On the (non-) existence of convex, calibrated surrogate losses for ranking. Advances in Neural Information Processing Systems, 25, 2012.

W. Chang, H. Yu, K. Zhong, Y. Yang, and I. S. Dhillon. Taming pretrained transformers for extreme multi-label text classification. In R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, editors, KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, pages 3163–3171. ACM, 2020.

W.-C. Chang, D. Jiang, H.-F. Yu, C.-H. Teo, J. Zhang, K. Zhong, K. Kolluri, Q. Hu, N. Shandilya, V. Ievgrafov, J. Singh, and I. S. Dhillon. Extreme multi-label learning for semantic matching in product search. In Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2021.

E. Chien, J. Zhang, C.-J. Hsieh, J.-Y. Jiang, W.-C. Chang, O. Milenkovic, and H.-F. Yu. PINA: Leveraging side information in extreme multi-label classification via predicted instance neighborhood aggregation. arXiv preprint arXiv:2305.12349, 2023.

A. E. Choromanska and J. Langford. Logarithmic time online multiclass prediction. In Advances in Neural Information Processing Systems 28, pages 55–63. Curran Associates, Inc., 2015.

P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In Proceedings of the fourth ACM conference on Recommender systems, pages 39–46, 2010.

K. Dahiya, A. Agarwal, D. Saini, K. Gururaj, J. Jiao, A. Singh, S. Agarwal, P. Kar, and M. Varma. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In International Conference on Machine Learning, pages 2330–2340. PMLR, 2021.

K. Dahiya, N. Gupta, D. Saini, A. Soni, Y. Wang, K. Dave, J. Jiao, G. K, P. Dey, A. Singh, et al. NGAME: Negative mining-aware mini-batching for extreme classification. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pages 258–266, 2023a.

K. Dahiya, S. Yadav, S. Sondhi, D. Saini, S. Mehta, J. Jiao, S. Agarwal, P. Kar, and M. Varma. Deep encoders with auxiliary parameters for extreme classification. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 358–367, 2023b.

O. Dekel and O. Shamir. Multiclass-multilabel classification with more classes than examples. In Proceedings of the Thirteenth International Conference on

Artificial Intelligence and Statistics, volume 9 of JMLR Proceedings, pages 137–144, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

K. Dembczyński, W. Kotłowski, O. Koyejo, and N. Natarajan. Consistency analysis for binary classification revisited. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 961–969, International Convention Centre, Sydney, Australia, 2017. PMLR.

K. J. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. An exact algorithm for f-measure maximization. In Advances in Neural Information Processing Systems 24, pages 1404–1412. Curran Associates, Inc., 2011.

J. Deng, S. Satheesh, A. C. Berg, and F. Li. Fast and balanced: Efficient label tree learning for large scale object recognition. In Advances in Neural Information Processing Systems 24, pages 567–575. Curran Associates, Inc., 2011.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.

J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In Proceedings of the 27th International Conference on International Conference on Machine Learning, page 327–334, 2010.

I. Evron, E. Moroshko, and K. Crammer. Efficient loss-based decoding on graphs for extreme classification. Advances in Neural Information Processing Systems, 31, 2018.

M. Frank and P. Wolfe. An algorithm for quadratic programming. Naval Research Logistics Quarterly, 3(1-2):95–110, 1956. doi: 10.1002/nav.3800030109.

D. Gillick, A. Presta, and G. S. Tomar. End-to-end retrieval in continuous space. arXiv preprint arXiv:1811.08008, 2018.

C. Guo, A. Mousavi, X. Wu, D. N. Holtmann-Rice, S. Kale, S. Reddi, and S. Kumar. Breaking the glass ceiling for embedding-based classifiers for large output spaces. Advances in Neural Information Processing Systems, 32, 2019.

N. Gupta, D. Khatri, A. S. Rawat, S. Bhojanapalli, P. Jain, and I. Dhillon. Dual-encoders for extreme multi-label classification. arXiv preprint arXiv:2310.10636, 2023.

S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, KDD '16, page 935–944, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939756.

H. Jain, V. Balasubramanian, B. Chunduri, and M. Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pages 528–536, Melbourne VIC Australia, Jan. 2019. ACM. ISBN 978-1-4503-5940-5. doi: 10.1145/3289600.3290979.

K. Jasinska and K. Dembczyński. Bayes optimal prediction for ndcg@k in extreme amulti-label classification. In From Multiple Criteria Decision Aid to Preference Learning Workshop, 2018.

K. Jasinska and N. Karampatziakis. Log-time and log-space extreme classification. CoRR, abs/1611.01964, 2016.

K. Jasinska, K. Dembczyński, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, and E. Hullermeier. Extreme f-measure maximization using sparse probability estimates. In Proceedings of The 33rd International Conference on Machine Learning, volume 48 of JMLR Workshop and Conference Proceedings, pages 1435–1444, New York, USA, 2016. PMLR.

K. Jasinska-Kobus, M. Wydmuch, K. Dembczynski, M. Kuznetsov, and R. Busa-Fekete. Probabilistic label trees for extreme multi-label classification. arXiv preprint arXiv:2009.11218, 2020.

S. Jayaram Subramanya, F. Devvrit, H. V. Simhadri, R. Krishnawamy, and R. Kadekodi. Diskann: Fast accurate billion-point nearest neighbor search on a single node. Advances in neural information processing Systems, 32, 2019.

T. Joachims. A support vector method for multivariate performance measures. In Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), August 7-11, 2005, Bonn, Germany, 2005.

P. Kar, H. Narasimhan, and P. Jain. Surrogate functions for maximizing precision at the top. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 189–198, 2015.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 2564–2572. PMLR, 10–15 Jul 2018.

S. Khandagale, H. Xiao, and R. Babbar. Bonsai: diverse and shallow trees for extreme multi-label classification. Machine Learning, 109(11):2099–2119, 2020.

S. Kharbanda, A. Banerjee, E. Schultheis, and R. Babbar. CascadeXML: Rethinking transformers for end-to-end multi-resolution training in extreme multi-label

classification. Advances in Neural Information Processing Systems, 35:2074–2087, 2022a.

S. Kharbanda, D. Gupta, E. Schultheis, A. Banerjee, V. Verma, and R. Babbar. Gandalf: Data augmentation is all you need for extreme classification. 2022b.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

W. Kotłowski and K. Dembczyński. Surrogate regret bounds for generalized classification performance metrics. Machine Learning, 10(4):549–572, 2017.

O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent binary classification with generalized performance metrics. In Advances in Neural Information Processing Systems 27, pages 2744–2752. Curran Associates, Inc., 2014.

O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent multilabel classification. In Advances in Neural Information Processing Systems 28, pages 3321–3329. Curran Associates, Inc., 2015.

J. Leskovec, A. Rajaraman, and J. D. Ullman. Mining of Massive Datasets. Cambridge University Press, USA, 2nd edition, 2014. ISBN 1107077230.

Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017.

Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence, 42(4):824–836, 2018.

T. K. R. Medini, Q. Huang, Y. Wang, V. Mohan, and A. Shrivastava. Extreme classification in log memory using count-min sketch: A case study of amazon search with 50m products. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 13265–13275. Curran Associates, Inc., 2019.

A. K. Menon, A. S. Rawat, S. Reddi, and S. Kumar. Multilabel reductions: what is my loss optimising? In Advances in Neural Information Processing Systems 32, pages 10600–10611. Curran Associates, Inc., 2019.

P. Mineiro and N. Karampatziakis. Fast label embeddings via randomized linear algebra. In Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases - volume Part I, volume 9284 of Lecture Notes in Computer Science, page 37–51, Gewerbestrasse 11 CH-6330, Cham (ZG), CHE, 2015. Springer.

A. Mittal, N. Sachdeva, S. Agrawal, S. Agarwal, P. Kar, and M. Varma. ECLARE: Extreme classification with label graph correlations. In Proceedings of the Web Conference 2021, pages 3721–3732, 2021.

H. Narasimhan, H. Ramaswamy, A. Saha, and S. Agarwal. Consistent multiclass algorithms for complex performance measures. In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2398–2407, Lille, France, 07–09 Jul 2015. PMLR.

H. Narasimhan, H. G. Ramaswamy, S. K. Tavker, D. Khurana, P. Netrapalli, and S. Agarwal. Consistent multiclass algorithms for complex metrics and constraints. 2022. doi: 10.48550/ARXIV.2210.09695.

N. Natarajan, O. Koyejo, P. Ravikumar, and I. Dhillon. Optimal classification with multivariate losses. In Proceedings of The 33rd International Conference on Machine Learning (ICML), pages 1530–1538, 2016.

A. Niculescu-Mizil and E. Abbasnejad. Label filters for large scale multilabel classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, volume 54 of Proceedings of Machine Learning Research, pages 1448–1457, Fort Lauderdale, FL, USA, 2017. PMLR.

J. Pearl. Heuristics: intelligent search strategies for computer problem solving. Addison-Wesley Longman Publishing Co., Inc., USA, 1984. ISBN 0201055945.

Y. Prabhu and M. Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, page 263–272, New York, NY, USA, 2014. Association for Computing Machinery.

Y. Prabhu, A. Kag, S. Gopinath, K. Dahiya, S. Harsola, R. Agrawal, and M. Varma. 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, pages 441–449, 2018a.

Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, and M. Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In Proceedings of the 2018 World Wide Web Conference, page 993–1002, Republic and Canton of Geneva, CHE, 2018b. International World Wide Web Conferences Steering Committee.

M. Qaraei, E. Schultheis, P. Gupta, and R. Babbar. Convex surrogates for unbiased loss functions in extreme classification with missing labels. In Proceedings of The Web Conference 2021, WWW '21, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3442381.3450139.

H. G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option. Electronic Journal of Statistics, 12(1):530 – 554, 2018. doi: 10.1214/17-EJS1388.

P. Ravikumar, A. Tewari, and E. Yang. On NDCG consistency of listwise ranking methods. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 618–626. JMLR Workshop and Conference Proceedings, 2011.

D. Saini, A. K. Jain, K. Dave, J. Jiao, A. Singh, R. Zhang, and M. Varma. Galaxc: Graph neural networks with labelwise attention for extreme classification. In Proceedings of the Web Conference 2021, pages 3733–3744, 2021.

Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata. Unbiased recommender learning from missing-not-at-random implicit feedback. In Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20, page 501–509, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368223. doi: 10.1145/3336191.3371783.

E. Schultheis, M. Wydmuch, R. Babbar, and K. Dembczyński. On missing labels, long-tails and propensities in extreme multi-label classification. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, page 1547–1557, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539466.

A. Shrivastava and P. Li. Improved asymmetric locality sensitive hashing (alsh) for maximum inner product search (mips). In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, page 812–821, Arlington, Virginia, USA, 2015. AUAI Press.

L. Song, P. Pan, K. Zhao, H. Yang, Y. Chen, Y. Zhang, Y. Xu, and R. Jin. Large-scale training system for 100-million classification at Alibaba. KDD '20, page 2909–2930, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403342.

Y. Tagami. AnnexML: Approximate nearest neighbor search for extreme multi-label classification. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 455–464, 2017.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. Journal of Machine Learning Research, 8(5), 2007.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

S. Vijayanarasimhan, J. Shlens, R. Monga, and J. Yagnik. Deep networks with large output spaces. CoRR, abs/1412.7479, 2014.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(12), 2010.

W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, and E. Hüllermeier. On the bayes-optimality of f-measure maximizers. Journal of Machine Learning Research, 15(1):3513–3568, 2014.

B. Wang, L. Chen, W. Sun, K. Qin, K. Li, and H. Zhou. Ranking-based autoencoder for extreme multi-label classification. arXiv preprint arXiv:1904.05937, 2019a.

X. Wang, R. Li, B. Yan, and O. Koyejo. Consistent classification with generalized metrics. arXiv preprint arXiv:1908.09057, 2019b.

T. Wei and Y.-F. Li. Does tail label help for large-scale multi-label learning? IEEE Transactions on Neural Networks and Learning Systems, 31(7):2315–2324, 2020. doi: 10.1109/TNNLS.2019.2935143.

T. Wei, W.-W. Tu, Y.-F. Li, and G.-P. Yang. Towards robust prediction on tail labels. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 1812–1820, 2021.

J. Weston, A. Makadia, and H. Yee. Label partitioning for sublinear ranking. In Proceedings of the 30th International Conference on Machine Learning, volume 28 of JMLR Workshop and Conference Proceedings, pages 181–189, Atlanta, Georgia, USA, 2013. PMLR.

M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczyński. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31, pages 6358–6368. Curran Associates, Inc., 2018.

F. Yang and S. Koyejo. On the consistency of top-k surrogate losses. In International Conference on Machine Learning, pages 10727–10735. PMLR, 2020.

H. Ye, Z. Chen, D.-H. Wang, and B. Davison. Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification. In International Conference on Machine Learning, pages 10809–10819. PMLR, 2020.

N. Ye, K. M. A. Chai, W. S. Lee, and H. L. Chieu. Optimizing f-measures: A tale of two approaches. In Proceedings of the 29th International Coference on International Conference on Machine Learning, page 1555–1562, Madison, WI, USA, 2012. Omnipress.

C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang. Learning deep latent space for multi-label classification. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

I. E. Yen, X. Huang, W. Dai, P. Ravikumar, I. Dhillon, and E. Xing. PPDsparse: A parallel primal-dual sparse method for extreme classification. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, page 545–553. Association for Computing Machinery, 2017.

R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in neural information processing systems, 32, 2019.

H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In Proceedings of the 31st International Conference on Machine Learning, volume 32 of JMLR Workshop and Conference Proceedings, pages 593–601, Bejing, China, 2014. PMLR.

H.-F. Yu, K. Zhong, J. Zhang, W.-C. Chang, and I. S. Dhillon. Pecos: Prediction for enormous and correlated output spaces. Journal of Machine Learning Research, 23(98):1–32, 2022.

M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. J. Mach. Learn. Res., 11:111–130, mar 2010. ISSN 1532-4435.

J. Zhang, W.-c. Chang, H.-f. Yu, and I. Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. Advances in Neural Information Processing Systems, 34, 2021.

R. Zhang, Y.-S. Wang, Y. Yang, D. Yu, T. Vu, and L. Lei. Long-tailed extreme multi-label text classification with generated pseudo label descriptions. arXiv preprint arXiv:2204.00958, 2022.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5(Oct):1225–1251, 2004.

W. Zhang, J. Yan, X. Wang, and H. Zha. Deep extreme multi-label learning. In Proceedings of the 2018 ACM on international conference on multimedia retrieval, pages 100–107, 2018.

J. Zhuo, Z. Xu, W. Dai, H. Zhu, H. Li, J. Xu, and K. Gai. Learning optimal tree models under beam search. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, 2020. PMLR.

D. Zong and S. Sun. GNN-XML: graph neural networks for extreme multi-label text classification. arXiv preprint arXiv:2012.05860, 2020.