Prof. Johannes Fürnkranz
Johannes Kepler University Linz
Institute for Application-Oriented Knowledge Processing (FAW)
Altenberger Straße 66B
4020 Linz
Austria
E-mail: juffi@faw.jku.at

Linz,9.8.2025

Reviewer's opinion
on Ph.D. dissertation authored by
Marek Wydmuch
entitled:
Addressing the long-tail problem
in extreme multi-label classification

1. Problem and its impact

Multi-label classification has become a standard problem in Machine Learning. Whereas the standard supervised learning setting is classification, which is a mapping of the input to one unique output label, multi-label classification (MLC) generalizes this setting by allowing a mapping to a (sub-)set of all possible labels. Practical examples of this problem are, e.g., assignments of (multiple) keywords to documents, or identifying objects in a given image.

Extreme multi-class classification (XMLC) denotes the challenge that in many MLC problems the number of possible labels may be extremely large, many of which will only be rarely observed in the training or test data. While, e.g., cats or dogs will frequently appear on training images, a platybus or an okapi will occur much more rarely. The frequencies of the occurrence of the possible labels thus has a long-tail distribution, i.e., most of the labels will only occur rarely. An extreme multi-label prediction problem of high practical relevance is, e.g., at the core of recommender systems, where a high number of products can be predicted as being of interest to a given user.

This dissertation argues that in such settings the prediction of rare long-tail labels is often very important (e.g., in a book recommendation setting, the successful recommendation of a rarely bought book has higher value than the prediction of the current bestseller), but purely reflected in conventional evaluation measures that are commonly used for measuring the performance of an algorithm on multi-label problems. Moreover, practical systems also often require to make a constant, fixed number of predictions (e.g., because the user interface of the recommender systems provides exactly k slots for advertising potentially interesting products to the customer).

This thesis theoretically analyzes this practically relevant setting in several variations, and complements the results with an empirical study on well-known real-world benchmark tasks. It thus increases our scientific understanding of this problem class, as well as delivers results that are of practical relevance.

2. Contribution

The main contribution of this work is a thorough and systematic theoretical investigation of a variety of different measures for evaluating MLC algorithms. Unlike conventional single-label classification, this problem has multiple dimensions, in that for every example a set of labels is predicted, and has to be compared to the ground truth – also a set of labels. First, there are many different ways for assessing the prediction performance on sets, recall and precison with their harmonic average F1 being the most prominent ones in the MLC context. Second, these predictions can be aggregated in different ways, e.g., by averaging the set prediction performance over all instances (instance-wise), or by first evaluating each label independently and then aggregating the performance over all labels (label-wise). Moreover, even the averages can be computed in different ways, e.g., by aggregating over all instances (micro-averaging) or first grouping according to labels and then average the label-wise performance (macro-averaging). One can also distinguish whether unset labels can be assumed to be false, or just missing, yielding propensity-based variants. A final distinction can also be made on the testing scenario, where the labels could be optimized for a fixed test set, as in the case when recommendations have to be made for an established set of customers (expected test set utility, ETU), or the predictive performance should be optimized over the entire joint distribution of customers and labels (population utility, PU). Prior to this thesis, there were a few isolated results on how to optimize some of these measures, this work gives a comprehensive treatment of this subject, which not only puts prior work into a coherent framework but also closes the difficult gaps, such as the optimization of macro-averaged measures.

Moreover, the thesis also proposes algorithms for optimizing the obtained measures, derives theoretical bounds on their expected performance, and empirically evaluates them on a variety of standard benchmark problems in the XMLC field. In Chapter 8, it also proposes algorithms for efficient methods for prediction under budget constraints. Personally, I found this to be particularly rewarding, even though not all the details became clear to me (e.g., in what way is BF* search more general than A* search, and what is the idea behind the heuristic function f(x,v) as defined in equation (8.7)? – the given explanation that it "estimates the gain of reaching the best label node in a subtree of node v" did not explain the nature of the defined estimate).

The results of this thesis have been published in the best venues of our field. I refrain deliberately from saying “some of the best”, because the list is impressively complete, encompassing ICML, NeurIPS, and ICLR, as the current three flagship conferences in machine learning, as well as KDD and SIGIR as the leading conferences in knowledge discovery in databases and information retrieval, respectively. All these conferences are very selective and highly competitive, so that already one or two contributions in these venues can be considered as an indicator for the quality of a thesis – it is certainly rare to cover all of them during a single PhD.

A grain of salt is maybe that all papers have been authored by 4-6 co-authors, where Mr. Wydmuch is first author on two, second author on five, and third author on one paper, so that the contribution of the candidate can not always be clearly identified. In particular for the contributions in chapters 6 and 7, which are key to this thesis, Mr. Wydmuch has only been second author of the underlying publications. However, in my opinion the coherence of the presentation of this work as a thesis clearly demonstrates a more than sufficient individual contribution, so that I do not have any reservations about the author’s achievements: they are, in my opinion, clearly above an average thesis in our field.

3. Correctness

The presentation of the work impresses with its formal rigor and elegance. It both sheds a new unifying light upon previous work in this area, as well as systematically fills in many of the gaps that are particularly relevant for XLMC. Also the experimental evaluation is very thorough and performed according to the state-of-the-art in this field. Overall, I think this work is an exemplary combination of theory and practice, lacking maybe only a concrete applied use case (as opposed to routine evaluation on benchmark data that are derived from real-world applications).

The reason why I think the latter could be of importance, is because I do think that the motivation behind this work, namely the importance of predicting infrequent long-tail labels, is not directly evaluated in the studied settings (which do follow the conventions for a good evaluation in this research area). While the author clearly and convincingly demonstrates that various loss functions which give higher weights to long-tail predictions can be optimized with the proposed algorithms, these measures only seem surrogates to measures that would be relevant for a real application, such as "increased sales on long-tail products". To this end, one could have also designed entirely different measures that primarily focus on low-frequency labels, such as the recall on the *c*% least frequent labels, or an inverse DCG measure where the weights are distributed according to the (true) occurrence frequency of labels. However, it is always easy to suggest additional work, the main criticism is not that these particular or other variants have not been investigated, but that the effect of the macro-averaged measures on the improved long-tail prediction could have been more clearly demonstrated.

Also the discussion of the results could at times have been a bit deeper. For example, in Table 1.1, which nicely illustrates that ignoring the tail labels will have a strong impact on macro-averaged measures such as Precision@k, this effect could have multiple reasons. When 80% tail labels are completely ignored, their label-wise recall will, for example, effectively be 0 on all examples, which means that 80% of the terms in the macro-averaged recall are 0. Clearly, with an increasing collective label mass on these examples, it becomes increasingly difficult to compensate this a priori disadvantage on long-tail labels with an increased recall on the 20% head labels. Similar arguments can be made for other macro-averaged measures, such as precision or F1. So, while we clearly see that completely ignoring these labels ("as they would not exist") does have a negative impact, it is not so clear to what extent the optimization improves tail label prediction.

Another case where I would have appreciated a somewhat more elaborate discussion is with the introduction of cp-Lipschitz functions, where, contrary to to regular Lipschitz bounds, the Lipschitz constant may depend on the number of set labels cp. It did not become clear to me, why this notion is necessary. Wouldn't it suffice to, e.g., assume single Lipschitz constant which is an upper bound (e.g., assuming that all labels are set for all examples, or the max over all example-dependent constants?). This is not meant to criticize the validity of the theoretical results, it is just an illustration that the presentation could sometimes be a bit more elaborate, in particular as machine learning research is traditionally more an algorithmic than a mathematical field (even if this has shifted considerably in recent years).

A very nice example can, on the other hand, be found in Setion 5.4., where the effect of the budget constraint (i.e., the requirement to predict exactly k labels) prevents an independent treatment of the

labels in Hamming and Jaccard scores, which could be optimized in that way in an unconstrained setting. More examples of this kind, also demonstrating clear practical advantages in application settings (beyond a decrease in the loss functions that are optimized) would have added to the thesis.

Overall, however, I would like to point out that the above criticisms are "luxury problems". The thesis is at an exceptionally high level, I would not hesitate to put it into the 5% best theses that I have seen in my career.

4. Knowledge of the candidate

The first two chapters of this thesis are introductory material, providing a general introduction into the research questions studied in this work and to the field of extreme multi-label classification. While the material presented in Chapter 2 is naturally well-known, the candidate nevertheless develops a very independent exposition, clearly tailored to and preparing the reader for the material in the later chapters. However, already in Chapter 1, we can find interesting original material in the form of the results of an extensive experimental study (Table 1.1) which demonstrates that the effect of simply ignoring long-tail levels is particularly strong on macro-averaged measures, the main topic of this work (as discussed above).

In the course of the thesis, the material gradually blends from known to original material. Chapter 3, e.g., presents standard instance-wise evaluation metrics for XMLC, but already provides an, as far as I could tell, original derivation of the form of the optimal classifier for these metrics and its regret bounds. Chapters 4 – 5 also mix previous work with new results. Chapters 6 – 8, along with the empirical evaluation in Chapter 9 form the key contributions of this work, and consists almost exclusively of novel results (first presented in the underlying publications).

For me, it is clear that the author has a very deep knowledge in the field of extreme multi-label classification. The more than 150 references cited in this work include (to my knowledge) all relevant works in this area.

5. Other remarks

The thesis is written in excellent English, and easy to read. There are a few remaining mistakes in grammar (e.g., missing or surplus articles), typos (e.g., a nice one is "nose" instead of "noise" on p.11), or incoherent capitalization (in particular in the bibliography), but they are certainly within the limits of what can be reasonably expected (probably much less than in this referee report). Particularly laudable is the extremely clear and coherent formal notation of this work, nicely summarized in a section at the beginning of the thesis. Proof ideas are sufficiently well sketched in the main text, detailed proofs are collected in an appendix, which increases the readability of the work substantially.

The results are presented in illustrative graphs and tables, which clearly highlight the key findings, while at the same time showing the full breadth of the performed experimental evaluation. A minor point of criticism is that, in my opinion, the authors sometimes relies a bit too much on the formal presentation of the results, as has also been mentioned in several instances above. The thesis could sometimes gain from more motivating examples as well as an accompanying less formal description of the intuitions behind the ideas. But these can be found in some cases, such as the motivation for the example- and label-dependence of propensities in Section 4.2.1, or the demonstration of the effect of budget-constrained optimization in 5.4. More of that would have further increased the readability.

Also, the various dimensions of the problem of defining XMLC evaluation measures (as discussed above in Section 2) could have been complemented with straight-forward graphical illustrations, which would have been useful for potential readers that are new to this area.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the A*ct of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

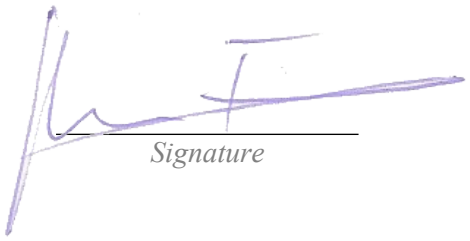
B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**?[2]

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

Moreover, taking into account its very mature formal rigor, the exemplary way in which it combines a coherent theoretical exposition with a strong experimental evaluation, and the exceptionally high quality of the venues in which the results of the thesis have been published, I **recommend to distinguish** the dissertation for its quality.



Signature

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276

[2] The term “Information and Communication Technology” was provided in the form that I used as a basis for this review. I answer this affirmatively in the sense that I am 100% convinced that the candidate has the general theoretical knowledge and understanding in the somewhat more narrow area of “Machine Learning”, for which I feel qualified to make such an assessment.