Wrocław, November 26th, 2021

Tomasz Kajdanowicz
Department of Artificial Intelligence
Faculty of Information and Communication Technologies
Wroclaw University of Science and Technologies
tomasz.kajdanowicz@pwr.edu.pl

Reviewer's opinion
on Ph.D. dissertation authored by
Kamila Jasińska-Kobus
entitled:
Efficient algorithms for extreme multi-label classification
(pol. Efektywne algorytmy dla wieloetykietowej klasyfikacji ekstremalnej)

1 Introduction

The dissertation of Kamila Jasińska-Kobus contains 146 numbered pages, 11 main and 3 supplementary chapters. The introduction chapter (1) provides a general overview of extreme multi-label classification, related work, motivation, aim, scope and dissertation outline accompanied with the author's contribution. The following chapters provide (2) theoretical background, (3) an introduction to classification quality metrics. Then, the main contribution of the dissertation is presented, i.e., the proposal of probabilistic label trees (PLT) for extreme multi-label classification (4), their statistical (5), and computational complexity (6) analysis followed by the online version of PLT (7). The dissertation also contains implementation details (8), results obtained in the empirical evaluation of PLTs (9), and a discussion on open research directions (10), followed by a summary (11). After the considerable number of references related to the topic, three supplementary chapters are devoted to proofs omitted in the main thesis, information on the experimental setup, and notation. There is also a meaningful summary of the dissertation in Polish at the end.

2 Problem and its impact

The dissertation is placed in machine learning, and more specifically, it concerns the problem of classification and supervised learning. It considers the problem of extreme multi-label classification. Such a problem is solved by constructing a classifier with a learning algorithm that assigns subsets of predefined labels to objects of interest. Unlike the classic multi-label classification, its extreme version makes assignments within a vast pool of possible labels. Since several labels can be assigned to the observation simultaneously, the distribution of labels plays a vital role in designing the learning algorithm. Their proper modeling is fundamental as extreme multi-label classification may result in computational efficiency problems. Therefore there is a strong need to develop the methods that operate in-between the decomposition to binary classification per label and 1-vs-all multi-class label subset codding. Moreover, the considered problem exposes the additional difficulty of learning from a few instances of training observations for long-tail labels.

The problem to be dealt with in the dissertation is derived from a common-sense observation that optimal decisions in extreme multi-label classification can be determined by modeling the conditional probabilities of labels. However, the estimation of the conditional probabilities and resulting inference can be challenging and complex. Thus, the proposed organization of labels as a tree in which each label corresponds to one and only one path from the root to a leaf, is a very reasonable scheme. The considered problem of accurate and efficient learning and inference in extreme multi-label tasks by introducing label trees, extends the body of knowledge and smartly adopts experiences from application to related multi-class, hierarchical softmax problems [1] or multi-output regression (conditional probability trees) [2]. Based on the above mentioned, the dissertation hypothesizes that «there exists a class of statistically consistent learning algorithms for extreme multi-label classification whose computational complexity is sublinear with the number of labels». The problem to be tackled in the research emerged naturally and inexorably from the literature review. It was a result of critical analysis of previous 1-vs-all works that achieved hard-to-beat results, e.g., [3, 4, 5, 6], but having significantly longer prediction time than other methods. Mrs Jasińska-Kobus taking up the problem of providing competitive execution time and preserving high accuracy methods for extreme multi-label classification, set a bar very high for herself.

In summary, the Ph.D. candidate tackled an exciting and essential topic of effective learning and inference in the extreme multi-label classification task. The problem is thoroughly scientific and possesses practical meaning.

3 Contribution

The primary and original contribution of the dissertation is related to the idea of Probabilistic Label Trees (PLTs). The Ph.D. candidate aims at proposing an efficient factorization method for joint label space probability conditioned on input space. The proposal directly applies a chain rule along the paths in the factorization tree. Each intermediate node in the tree represents a subset of labels and leafs– particular single-labels. The general idea of learning and inference in this setting reduces the original multi-label problem to a number of binary classification problems. Such solutions for the multi-label domain is a straightforward adaptation of the learning reduction framework originally proposed for multi-class in [7], which is honestly mentioned in the dissertation.

What must be emphasized is the fact that the proposed extreme multi-label problem operationalization through PLTs is introduced, followed, justified, and proved using very well utilized basic concepts that lead through the dissertation fully navigated. An adequately dosed introduction to theoretical background, seen from statistical decision theory point of view, firstly defines the risk of a classifier, the Bayes optimal classifier, and the regret of a classifier. Then 0/1 loss in binary classification is ported to present the importance of estimation of conditional probabilities of labels and the need of operating with L_1 estimation error (for strongly proper composite loss functions allowing even to bound the L_1 error). Then, the consistency of classifiers is defined in the context of surrogate losses and learning reductions, which reveals the prospective mechanics of PLTs.

The dissertation briefly overviews the most popular extreme multi-label classification metrics with the inclination to the forms of the corresponding optimal classifier. The author concisely shows for which metrics the optimal decisions can be determined through the conditional probabilities of labels. A substantial part of the presented consideration is an original contribution of the author: for precision@k and the pick-one-label heuristic [8, 9] and NDCG@k [10]. It was shown that popular precision@k, DCG@k, Hamming loss, and the micro- and macro F-measures could be optimally determined through the conditional probabilities of labels. This result is so significant that it makes possible assumptions that «estimating the conditional probability of labels, and then predicting the labels according to the form of the Bayes optimal classifier, is a consistent strategy for solving extreme multi-label classification under those metrics.»

Based on such a well-developed foreground Ph.D. candidate presents the main contribution after her previous publications [11, 8, 9] - the Probabilistic Label Trees (PLTs), a mechanism for efficient estimation of the conditional probabilities of labels. It contains a bit condensed description of the probabilities factorization used by PLTs. There is proposed a training procedure of PLTs (Algorithm 1 and 2) that learns probabilistic classifier for each node in the tree using subsets of observations assigned with particular subsets of labels. Inference, that estimates the conditional probability of a particular label, is then the product of probability estimates of these classifiers on the path from root to leaf. The author discusses several prediction procedures that allow obtaining labels rather than their distribution: threshold-based predictions (Algorithm 3), which assign labels if the probability estimate is greater or equal certain threshold, prediction of top k labels with the highest probability estimates (Algorithm 4) and its alternative beam search algorithm that allows controlling the uniform-cost of search and traverses more wisely the tree.

The dissertation contains a solid statistical analysis of PLTs. The reported analytical results show the upper bound of L_1 estimation error of conditional probabilities of labels by means of the L_1 error of the node classifiers. There is a successful attempt to generalize this result to a broad class of strongly proper composite losses – L_1 error of conditional probability estimates is straightforwardly connected with a learning algorithm used in the tree nodes. Then the author applies all obtained bounds to demonstrate the regret bounds for generalized classification performance metrics. It turns out that «for conditional probability estimates, there exists a vector thresholds, for which the regret of a generalized performance metric is upper-bounded solely by regrets of node classifiers, expressed in terms of a strongly proper composite loss function.» There are also analyzed the precision@k and DCG@k metrics that use the previously obtained bounds for strongly proper composite loss functions followed by a note that PLTs are strictly related to hierarchical softmax [1]. Altogether, the dissertation offers a bloated collection of corollaries, theorems, and proofs giving insight into the statistical properties of the proposed methods and, more importantly, high maturity and a great understanding of the domain.

The analysis of the computational complexity of PLTs training and prediction is an another contribution of the dissertation. The analysis is delivered through training and prediction costs expressed in the number of updated or evaluated node classifiers and then bounded to relate the cost of prediction with the cost of training and the L_1 error of the classifiers.

The dissertation also presents proposals for the online versions of probabilistic label trees. This version's method aims to train node classifiers online without prior knowledge of the set of labels. Mrs. Jasińska-Kodus introduced the extension to PLTs that goes beyond batch mode learning to incremental, way with an assumed tree structure (Algorithm 6) and without knowing the tree structure, built as-the-data-arrives (Algorithms 7-12). The approach is based on keeping (adding, updating) an appropriate set of classifiers in the tree ensemble. The crucial part of the tree updating strategy is related to the policy method that decides which of the three variants to follow and selects a particular node to start the extension from. The proposed strategy is a trade-off choice from the possible policies capable of factorizing the joint label probability well enough and efficiently. This part of the dissertation is concluded with a proofed theorem that the proposed online PLT is proper and efficient.

The author of the dissertation also provides an overview of the possibilities of PLTs' implementation. We can find a compact description of the existing packages that implement various approaches to extreme multi-label classification and data representation. The description equips the reader with implementation direction across training, prediction, dense and sparse representation, tree construction, or ensembles of PLTs.

The described algorithms and theoretical findings were carefully evaluated with a comprehensive empirical study. Among others, there are reported experimental results for:

(1) different design choices of PLTs

For the (1a) batch and incremental learning of node classifiers (with logistic and squared hinge loss), the author observes that «none of the configurations strictly dominates the others». However, the author also claims that «the batch training with squared hinge loss is the most reliable, without outlying results». The reviewer is not equipped with the details or methodological explanation of these conclusions (e.g., statistical tests). Studying two approaches to prediction methods (1b), uniform-cost search based or beam search based, in online and batch implementations «both methods perform very similarly in terms of *precision@k*». While considering (1c) training and prediction with sparse and dense representation, it is shown that «sparse representation achieves higher *precision@k*» than dense. Different tree-building strategies (1d) are also tested. The results show that it is justified to build meaningful label probability factorization over the tree instead of randomly distributing labels in it. Closing experiment for the design on PLTs checks if they work in ensembles (1e). It seems that ensembles provide better accuracy, and squared hinge loss gains slightly more than the ensembles trained with logistic loss.

(2) evaluation in terms of Hamming loss, micro- and macro-F measure that verifies the generalized performance metrics according to the contributed theoretical part of the dissertation

The experimental procedure implemented the decision thresholds suited for a particular metric as obtained in the theoretical analysis. It was shown that following the analytical expectations reveals significantly better results: threshold 0.5 performs best for Hamming loss, while the thresholds tuned with online F-measure optimization tend to outperform in F-1 measures.

(3) testing and finally confirming the suboptimality of hierarchical softmax with pick-one-label heuristic

Two experiments devoted to the comparison of hierarchical softmax with the pick-one-label heuristic (HSM-POL) and PLTs have shown that HSM-POL returns suboptimal sòlutions concerning *precision@k*. It was shown by measuring the performance on synthetic and benchmark data sets. Here, it is worth emphasizing that from the well-thought-out experimental procedure, it can be directly observed that PLTs clearly outperform HSM-POL on data with conditionally dependent labels (confirmed by statistical test). It aligns with the contributed theoretical results.

(4) performance of the fully online variant of PLTs, in which both node classifiers and tree structure are built incrementally on a sequence of training observations

The obtained results show that online PLT with the tree built according to complete tree policy from Algorithm 9 is even competitive with batched PLT trained on a complete binary tree.

(5) comparison of PLTs to relevant state-of-the-art algorithms

Finally, the author of the dissertation presents the results of quality and computational performance comparison to the state-of-the-art algorithms. It can be noted that PLTs are competitive to the other approaches (in *precision@1*) on the majority of benchmark data sets. Moreover, PLTs are the fastest in almost all cases in training and prediction, having the smallest model sizes. Comparisons to other methods were performed with original implementations of all competitors.

All the experiments were performed using 8 medium to large size datasets from Data Extreme Classification Repository[1] and synthetic dataset, as mentioned above.

The dissertation also contains an insightful discussion on ideas related to PLTs that have not been published, on limitations of PLTs, and on several open research directions related to PLTs. To name a few, (1) BR-trees are presented [11], (2) additional possible extensions of tree building strategies (beyond random, k-means clustering, complete tree-building policy) are discussed, (3) relation to methods that bound the prediction cost by a function of the distance from the optimal prediction, or (4) limitations of PLTs not addressed, and that was out of scope of the dissertation, i.e., handling of rare labels, dealing with not observed labels, as well as (5) open research directions, i.e., making PLTs suitable for more performance metrics, usage of other probabilistic classifiers and other applications.

To summarize the contribution of the dissertation, the reviewer is firmly convinced to claim that **contribution is very extensive, original, exciting, and, above all, very thoroughly analyzed analytically and empirically**. The list of publications that constitute the backbone of this dissertation is the best confirmation that the author has made a significant contribution to the development of the extreme multi-label classification. The Ph.D. candidate published papers at the top-tier conferences, including International Conference on Machine Learning main conference [12] and workshop [11], Neural Information Processing Systems [13] as well as International Conference on Artificial Intelligence and Statistics [14]. Her papers were cited almost 200 times (Google Scholar).

In summary, the Ph.D. candidate proposed a novel theoretical and methodological slant on an extreme multi-label classification organized in PLTs and has created interesting analytical friction by bringing together hitherto expected by the community topics. It was a pleasure for the reviewer to follow the presented work.

4 Correctness

The content of the dissertation is presented with appropriate clarity. A skilled reader is able to keep up with a relatively complex analytical concepts flow, especially in the proofs of the theorems. There are logical and rational links between the dissertation's parts, which constitutes the intellectual

[1] unfortunately, the reviewer was not able to reach the repository at the link given in the dissertation, http://manikvarma.org/downloads/XC/XMLRepository.html

wholeness of the submission. The research and the written part of the dissertation is the candidate's own work, which, according to publication record, was performed as part of a more extensive research program in a team. The candidate demonstrated that she has detailed knowledge of original sources in the extreme multi-label classification and deeply understood all the theoretical and methodological issues related to the research. To the best knowledge of the reviewer, the dissertation related to all significant secondary literature sources available at the time the methods were proposed, and contained their critical assessment making succinct, penetrating, and challenging to read review.

In general, the candidate justified the chosen methodology, showing an appropriate rationale in each case. It applies to both analytical and experimental parts of the work. It was clearly demonstrated why each particular analysis was conducted, how the analysis was done, and what the analysis tells us about the gathered results. In most cases, there was a fully expressive discussion that summarised parts of introduced concepts without undue repetition. All the introduced concepts, e.g., PLTs or online PLTs, were linked to the literature review, and the empirical findings were interpreted and related to theoretical contribution.

All the profs presented in the main part and the appendix of the dissertation seem to be correct and consistent; they were repeated in part by the reviewer. The robust hypothesis of the thesis on the existence of learning algorithms for extreme multi-label classification whose computational complexity grows slower than linearly with the number of labels is met in chapter 6 (under assumptions regarding the tree structure). It has also been shown experimentally and repeated partially by the reviewer to confirm (using the detailed description in appendix B). The code of the methods and datasets (partially) are available on the web.

5 Knowledge of the candidate

The reviewer is fully convinced that the candidate has great general knowledge and understanding of the scientific discipline the dissertation was submitted to, i.e., Information and Communication Technology. It can be observed in all the chapters of the submission. It relates, among others, to machine learning, probability and statistics, algorithmics, computational complexity, optimization, data structures, as well as decision and learning theory. Reading the discussion part of the dissertation revealed to the reviewer that the candidate has learned a lot during the course of the work.

6 Other remarks

The critical remarks and comments provided below should be treated as a scientific discussion that aims at directing the author to more in-depth research in the field of extreme multi-label classification. Below the reviewer lists some minor remarks or questions that arose during the text review:

1. Proposition 4.1, the last equation should be $\eta_{v'}(\mathbf{x})$
2. Lemma 5.1 would look a bit more consistent with equation 4.2 (its explanation) if instead of $\eta_{\mathrm{pa}(r_T)}(\mathbf{x})$ there would be a similar distinction of root vs. other cases in text
3. There are few results in bold that were not the best results in Table 9.13

4. Does the order of appearance of new labels in an Online PLT, see Algorithm 9, affects the ability to generalize node classifiers (factorization of join label probability is performed depending on the arrival of new labels)? If the answer is yes, is it possible to propose a policy invariant to the order of new labels' arrival?
5. Would it be meaningful to provide statistical tests to experimental results comparisons?
6. In the online learning of the tree structure, one can expect to obtain the best possible factorization of the conditional probability of labels. In the setting, as presented in the dissertation, this is mainly accomplished by applying A_{policy}. It would be great to measure if the particular instance of the tree building policy builds optimal factorization. The reviewer would be delighted if the Ph.D. candidate could elaborate more on how to measure that.
7. Considering the experiment on PLT's generalization with sparse vs. dense representation, does this experiment examines the abilities of PLTs to work with sparse or dense representation rather than the impact of such representations on hierarchical clustering and generalization abilities of node classifiers?
8. The experimental part assumes some train/test splits as well as it is mentioned that «we focus on batch approaches which assume that labels are known». Please comment on how the PLT would act if the labels from the test set were not present in the training set.

7 Conclusion

Summing up the entire review, the reviewer wishes all doctoral students such scientific results, such research maturity, and such publications during theirs doctoral dissertation as well as careful and substantive supervision, as it was in the case of Mrs. Kamila Jasińska-Kobus.

Taking into account what I have presented above and the requirements imposed by Article 13 of the Act of March 14th 2003 of the Polish Parliament on the Academic Degrees and the Academic Title (with amendments), 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, and particularly the area of machine learning?

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 the high maturity of presented content and blameless publication record, the reviewer **recommend distinguishing the dissertation for its quality**.

Referencess

[1] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *International workshop on artificial intelligence and statistics*, pages 246–252, 2005.

[2] Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. In *International Conference on Algorithmic Learning Theory*, pages 247–262, 2009.

[3] Rohit Babbar and Bernhard 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, 2017.

[4] Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In *International conference on machine learning*, pages 3069–3077, 2016.

[5] Ian EH Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric 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*, pages 545–553, 2017.

[6] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8):1329–1351, 2019.

[7] Alina Beygelzimer, Hal Daumé, John Langford, and Paul Mineiro. Learning reductions that really work. *Proceedings of the IEEE*, 104(1):136–147, 2015.

[8] Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczyński. A no-regret generalization of hierarchical softmax to extreme multi-label classification. *arXiv preprint arXiv:1810.11671*, 2018.

[9] Kalina Jasinska-Kobus, Marek Wydmuch, Krzysztof Dembczynski, Mikhail Kuznetsov, and Róbert Busa-Fekete. Probabilistic label trees for extreme multi-label classification. *arXiv preprint arXiv:2009.11218*, 2020.

[10] Jasinska Kalina and Dembczynski Krzysztof. Bayes optimal prediction for ndcg@ k in extreme multi-label classification. In *Workshop on Multiple Criteria Decision Aid to Preference Learning (DA2PL)*, 2018.

[11] Kalina Jasinska and Krzysztof Dembczynski. Consistent label tree classifiers for extreme multi-label classification. In *Extreme Classification Workshop, International Conference on Machine Learning*, 2015.

[12] Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hullermeier. Extreme f-measure maximization using sparse probability estimates. In *International conference on machine learning*, pages 1435–1444, 2016.

[13] M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *Advances in Neural Information Processing Systems*, pages 6355–6366, 2018.

[14] Kalina Jasinska-Kobus, Marek Wydmuch, Devanathan Thiruvenkatachari, and Krzysztof Dembczynski. Online probabilistic label trees. In *International Conference on Artificial Intelligence and Statistics*, pages 1801–1809, 2021.

Kajdanowicz

dr hab. inż. Tomasz Kajdanowicz,
prof. PWr