Rohit Babbar,
Assistant Professor,
Department of Computer Science
Aalto University, Helsinki, Finland
tel : (+358)50 512 2646
rohit.babbar@aalto.fi

Aalto University

# Review of Kalina Kobus's Thesis : Efficient Algorithms for Extreme Multi-label Classification

## 1 Thesis summary

The broad area of this thesis is on the machine learning problem of multi-label classification in which the goal is to learn statistical models which can predict the subset of relevant labels from the set of all possible labels. In particular, this work attempts to address the particularly challenging setting where the number of possible labels are extremely large, of the order of Millions. Also referred to as Extreme multi-label classification, this framework has found numerous applications in web-search, recommendation systems and computational advertising. Due to the large-scale nature of the problem, efficient learning algorithms, preferably with statistical guarantees, are hence desirable.

In this thesis, the author proposes the Probabilistic Label Tree (PLT) model, which can be interpreted as a generalization of hierarchical softmax for multi-label classification. Not only does the PLT model provide a computationally efficient framework for extreme-scale classification problem, but its statistical properties are also thoroughly studied in this thesis.

In addition to the above contributions, concrete algorithmic implementations of the PLT model, in the form of ExtremeText and NapkinXC with open source codes, have been provided. These algorithms compare favourably to state-of-the-art methods on benchmark datasets in this domain. Furthermore, the versatile nature of these algorithms is demonstrated under various settings including (i) sparse and dense input data representations, (ii) Batch and online learning, and (iii) different tree-structures in the PLT architecture.

Apart from the introduction and motivation, the thesis is implicitly divided into three parts :

1. Theoretical framework - Multi-label learning, Performance metrics and their properties
2. Algorithmic framework - PLT model and their statistical and Computational aspects
3. Empirical evaluation - Concrete implementations of PLT model and relative comparisons

In the first part, covering chapters 2 and 3, the formal framework of binary, and multi-label classification is setup. In chapter 2, for the case of binary classification, the key notions of the Bayes optimal classifier, regret, and $L_1$ estimation error between the quantity of interest $\eta(x)$, the positive class probability for a given instance $x$, and its empirical estimate, are introduced. It is then argued that, under mild assumptions, $L_1$-error of a classifier can be upper bounded by its

regret on most commonly used convex loss functions. This idea is then extended to the multi-label setting for the case of Hamming loss case in which the loss decomposes over individual labels into separable binary problems. In chapter 3, first a discussion of generalized performance metrics is presented, followed by precision@k and recall@k, and their corresponding optimal strategies in terms of conditional probabilities $\eta_j(x)$ for label $j$. It is then shown that under conditional independence of labels, both lead to the same ordering of labels. Similar claims are proved for another set of metrics common in extreme classification literature, namely discounted cumulative gain (DCG@k) and its normalized variant (nDCG@k).

The algorithmic contribution of the thesis is covered in the second part, which spans over the chapters 4-to-7. In chapter 4, the hierarchical framework of PLTs is presented, and the formulas of conditional probabilities of labels at the leaf nodes in terms of the values for on internal tree-nodes is derived. The algorithmic procedures for training and prediction under the PLT model are discussed in detail. In particular, for the prediction part, two different schemes are discussed. The first one based on label-specific thresholds, which leads to optimal decision for many performance metrics. The second one is based on uniform cost-search and beam-search to predict the top-k labels with the highest estimated conditional probabilities $\hat{\eta}_j(x)$. In chapter 5, statistical analysis of PLT model is presented in three parts : (i) first the $L_1$ estimation error of the labels at the leaf nodes is bounded in terms of the same quantity for the internal tree-nodes on the path from the root to the leaf node, (ii) then this error is bounded in terms of a strongly proper composite loss function, and (iii) finally the previous results are then leveraged to obtain regret bounds on the generalized classification performance metrics as well as for precision@k and DCG@k. Also, it is demonstrated that the PLT model is a no-regret generalization to the famous hierarchical softmax model [7]. In chapter 6, the computational aspects of the PLT model are discussed. By defining the unit cost as a single access to a node classifier in the tree, it is shown that the training complexity of the model scales logarithmically in the number of labels. The expected prediction cost is then bounded in terms of the expected training cost. In chapter 7, based on the most recent work by Kalina, an online version of the PLT model is presented, which is useful in the real-world scenarios when the set of labels is not known in advance.

The third part of the thesis, in chapters 8 and 9, focusses on the concrete implementations of the PLT model and their comparison with state-of-the-art methods on benchmark datasets. In chapter 8, various design choices are presented, which include – batch vs online training of the node classifiers, sparse vs dense feature vectors, implications of uniform cost-search and beam-search in the prediction setting, configuration of the tree-structures, and the impact of an ensemble of PLT models. In chapter 9, first these above design are concretely evaluated against each other. In addition to the precision@k and nDCG@k metrics, computational complexity of training and prediction under the model is also discussed. Thereafter, PLT model is compared to state-of-the-art methods, namely FASTXML and PFASTREML as decision-tree based methods, and DISMEC and PPDSPARSE as smart one-vs-rest methods. It is demonstrated that the PLT models can give competitive predictive performance at a fraction of computational and storage costs.

Finally, the thesis concludes on an important note with discussions about open research directions and limitations of the PLT model. The main part of the thesis ends with a concise summary and related references. The appendix consists of two chapters, one with supplementary proofs and another with further details on the experimental setup. In terms of organisation, the thesis does an excellent job of appropriately separating the proofs which are somewhat dissociated from the main thread of the respective chapters by relegating them to the appendix.

## 2 General remarks

The thesis is mainly based on five papers co-authored by the PhD candidate. Two of these are journal papers, one of which is under review at Journal of Machine Learning Research, a top journal in this domain. The three conference papers have been published in the top-most venues in machine learning and includes, ICML 2016, NeurIPS 2018, and AISTATS 2020. In addition, few other papers have been published in workshops and open access repositories such as arxiv.org. The exact bibliographic details of the journal and conference articles are given below

- Journal papers

    1. K. Jasinska-Kobus, M. Wydmuch, K. Dembczynski, M. Kuznetsov, and R. Busa-Fekete. *Probabilistic Label Trees for Extreme Multi-label Classification*. In Journal of Machine Learning Research, under review, 2020
    2. K. Jasińska, K. Dembczyński, N. Karampatziakis. *Extreme classification under limited space and time budget*. In Schedae Informaticae. 2017

- Conference papers

    1. K. Jasinska, K. Dembczynski, 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, 2016
    2. 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 Advances in Neural Information Processing Systems, 2018
    3. K. Jasinska-Kobus, M. Wydmuch, D. Thiruvenkatachari, and K. Dembczyński .Online probabilistic label trees. In AISTATS 2020

In addition to the fundamental level of contributions made in this work as described above, this is also a very-written thesis. It does an excellent job of introducing various concepts, formal definitions, connection with existing works, and clearly differentiating the contributions from the previous research. Moreover, the organization of the manuscript is very clear in terms of motivation and introduction, theoretical framework, algorithmic contributions and empirical evaluation on standard benchmark datasets. Certainly, the work carried out as part of this thesis will be a model to emulate for ongoing and future research carried out by graduate students in this area.

## 3 Detailed comments

Below are some comments which could be addressed to further improve the coverage and presentation :

- Contributions in Chapter 3 - The preface of this chapter mentions about the contributions of thesis in the form of various papers by Kalina [4, 3, 8]. However, in the main part of the chapter, the exact contribution corresponding to these papers is somewhat limited, mainly on page 21. It would be nice to expand and refer to the above papers, co-authored by the PhD candidate, to the appropriate subsections of this chapter.

- Tail-labels and metrics - One of the limitation of the thesis is a limited discussion on tail-labels which are the infrequently occurring labels in the training/test data. Given that the focus and research interest on long-tail problems is growing in other areas (such as Computer Vision [1]) as well, it might help to discuss this key facet, which is so inherent to extreme classification setting, in the introduction section. Also, if available, it would be recommended to add results for the corresponding metrics (in the form of PSP@k and PSnDCG@k) which take into account the label propensities [2].

- Coverage of Deep learning methods - Of late, there has been a surge of deep learning methods in extreme multi-label classification. Some of these have not been covered in this thesis, such as those based on transformer models [5, 9]. It would further enrich the thesis and broaden its scope by adding these works in the related work section of chapter 1.

# 4 Minor remarks

- In terms of structure of various chapters, conclusion as a section is mentioned in Chapter 3 but not in other chapters, this should be made uniform,

- In table 2.1, it would nice to add the CPE loss formula and link function for squared hinge loss which is one of the most commonly used loss in the extreme classification setting,

- On page 12, where 0-1 loss function $\ell_{0/1}$ is defined, it should be mentioned that the prediction $h(x) \in \{0, 1\}$, as earlier in the beginning of the chapter it was defined differently $(h(x) \in \mathbb{R})$

- The reference for BONSAI TREE on page 4 is incorrect, also in the bibliography it could be referred to its official peer-reviewed version [6], which is updated one compared to the arxiv version,

- On page 34, it is mentioned 'The threshold-based prediction ... **in the next section**, ...', while there is no next section. I think what is referred here is the next paragraph.

# 5 Conclusion

The framework of Probabilistic Label Trees introduced as part of this thesis is a key contribution in the field of extreme multi-label classification. This is evident from its usage and adoption in a wide variety of algorithms in the contemporary literature. Apart from this algorithmic contribution, solid analysis on the statistical and computational fronts is also an integral part of this thesis.

I assess the current content of the thesis as excellent, and one of the top contemporary thesis compilation in this area. The comments above mainly in the form of minor remarks should be used for further improving its presentation. To summarise, it should be mentioned that the goal of the thesis as stated in the beginning has been achieved. Considering the above comments, I request for proceeding ahead to further stages of the doctoral examination of Kalina Kobus.

# 6 Evaluation

| Guiding questions: | Answers | | | | |
|---|---|---|---|---|---|
| | Definitely YES | Rather Yes | Hard to say | Rather no | Definitely NO |
| Does the dissertation present an original solution to a scientific problem? | X | | | | |
| After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of Information and Communication Technology ? | X | | | | |
| Does the dissertation support the claim that the candidate is able to conduct scientific work ? | X | | | | |

Moreover, taking into account *the excellence and impact of the scientific work* conducted during this thesis, I recommend to distinguish the dissertation for its quality.

Rohit Babbar, 09/06/2021

Rohit Babbar, PhD

# References


[1] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019.

[2] H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking and 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.

[3] K. Jasinska-Kobus and K. Dembczyński. Bayes optimal prediction for ndcg@k in extreme multi-label classification. In *From Multiple Criteria Decision Aid to Preference Learning Workshop*. 2018.

[4] K. Jasinska-Kobus, M. Wydmuch, K. Dembczyński, M. Kuznetsov, and R. Busa-Fekete. Probabilistic label trees for extreme multi-label classification. *CoRR*, abs/2009.11218, 2020.

[5] T. Jiang, D. Wang, L. Sun, H. Yang, Z. Zhao, and F. Zhuang. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. *arXiv preprint arXiv:2101.03305*, 2021.

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

[7] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005.

[8] 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 S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6355–6366. Curran Associates, Inc., 2018.

[9] 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.