Jan Mielniczuk, professor
Institute of Computer Science, PAS
&
Warsaw University of Technology

6.VIII.2025

Reviewer's report on Ph.D. dissertation authored by Marek Wydmuch entitled *Adressing the long-tail problem in extreme multi-label classification*

Thematic content of the thesis and its aim

The Ph.D. thesis by Marek Wydmuch (advised by prof. Krzysztof Dembczyński) is a thorough analysis of the risk (utility)-based approach to multi-label classification problems when the number of the labels is extremely large (Extreme Multilabel Classification, XMLC in short). The main aim of the thesis is to discuss three pivotal aspects of the problem: interplay between instance-based and macro-averaged utilities; properties of classifiers budgeted at k (@k classifiers); and consequences of using different types of utilities, in particular form and properties of corresponding optimal classifiers in family @k of classifiers. Specifically, the thesis considers three types of utilities: conditional (ETU), unconditional (EIU) and population utility (PU) as well as approximation of ETU called semi-empirical ETU. The main results concern establishing forms of the minimisers for different utilities in different classes of classifiers (@k classifiers, deterministic and randomised classifiers), regret analysis under various considered scenarios and practical consequences of the results (Block-Coordinate Ascent algorithm, Frank-Wolfe algorithm and estimation of posterior distribution using Probabilistic Label Trees). The theoretical results are supplemented by analysis of performance of studied classifiers on 9 real data sets conducted for considered utilities as well as analysis of optimisation of weighted utilities being convex combination of instance-wise and macro-averaged utilities.

Evaluation. General comments

The problems chosen as main topics of the thesis have obvious theoretical importance and their solutions, even partial, have far-reaching practical consequences. When the number of labels is very large, the unavoidable problem is occurrence of labels with very few positive instances which may make such labels under-privileged when instance-wise utilities are applied. This is especially important when the analysis of those labels corresponding to long tail of frequencies is considered vital. Thus the study of macro-averaged utilities which do not have this drawback is a natural one. Moreover, for this scenario, @k family of classifiers is a natural family of classifiers (actually, this

approach is analogous to k-best subset selectors in regression variable selection). Regret analysis is the state of the art method of checking how much one deviates in terms of utility/risk when the optimal theoretical solution is replaced by empirical utility maximiser.

Theoretical results in the thesis published in a series of papers co-authored by Candidate are novel and, to the best of my checking, formally correct. Theorems 3.2.2, 6.3.2, 6.6.1, 7.2.4 and 8.2.3 are in my view the most important theoretical results of the thesis. What I particularly like about the results are their practical implications and the way they are taken advantage of in the thesis: e.g. Theorem 6.3.2 concerning approximation of ETU utility by its semi-empirical version motivates Block-Coordinate Ascent (BCA) algorithm (Section 6.5) and Frank-Wolfe algorithm (Section 7.3) is underpinned by Theorem 7.2.4.

Theoretical results contained in the thesis are outcomes of collaboration of several researchers besides M. Wydmuch and his advisor prof. Krzysztof Dembczyński, thus the obvious question of Candidate's input arises. I have obtained from him and his advisor detailed description of the input (official thesis documents did not contain this important information) which I consider satisfactory. In particular, besides contributing to all chapters of the thesis, Candidate played an instrumental role in proposing and developing greedy BCA algorithm and developing Frank-Wolfe algorithm (which in my opinion belong to the highlights of the thesis) and proved the result (Theorem 6.7.1) on the regret of @k coverage which requires separate treatment as it is not *cp*-Lipschitz. Also, real-data analysis in Chapter 9 is his solo contribution to the content of the thesis.

The real-data analysis of Chapter 9 is elegantly divided into two main parts: the first which eliminates the influence of estimation error of multivariate posterior η (this is achieved by creating pseudo-empirical data via generating artificial labels from estimated $\hat{\eta}$) and the second part, for which η is unknown and estimated. The results of the first part confirm the theoretical results showing that the considered methods work best on the utilities which they aim to maximise. For the second part the best results were obtained for macro-averaged methods when they are optimised using BCA algorithm.

The lengthy thesis, which covers important an problem in Machine Learning, is self-contained and clearly written in English; the studied problems are well referenced and references reflect state of the art.

The thesis, based on papers co-authored by the Candidate (dating from 2018 on) is, as already said, voluminous and contains many significant results. Still, one may wonder whether less material (e.g. omitting critique of Jain's (2016) approach) would not result in more focused dissertation. On

the other hand, this would possibly make numerical analysis poorer as not all entities appearing there would be then discussed in depth.
The discussion of three problems main problems in the thesis mentioned above gives rises to some interesting issues. Let me state a few.
(i) It follows from Chapter 9 that estimation of multivariate posterior η has a major influence on properties of inferential procedures. This is also supported by several results on regret in the thesis which include L^1 or L^2 discrepancy between $\hat{\eta}$ and η in the bounds. The Author uses a specific estimator of η based on Probabilistic Label Trees. What is the rationale of using this specific estimate and how results would change when using other existing estimates of η?
(ii) The method applied in the paper is to reduce XMLC to a problem of smaller dimensionality k' and then to apply proposed procedures of large multi-label problem with k' categories. Thus perhaps the issue how to solve optimally large (but not extremely large) multi-label problem is worth pursuing;
(iii) Much attention is devoted to solving $@k$ problems e.g. for precision$@k$ or recall $@k$. As k is obviously unknown, perhaps considering k as a random parameter and finding optimal solutions for its specific distribution would help to circumvent a problem of focusing on a specific value of k. If intuition from variable selection applies here, this would result in a specific penalty added to the utility;
Specific comments
1. Regrets considered in the dissertation are considered conditionally on the data on which classifier is learnt. This is not sufficiently well stressed;
2. Sometimes, the discussion how the obtained results relate to known results are missing. In particular, there is no discussion of obtained results on missing data in relation to previous ones for $m = 1$. The conditional expectation for the utility derived in Theorem 4.1.2 coincides with the negative conditional risk on p. 23 in Bekker and Davis (2022) in the case of accuracy ($g_{fn} = g_{fp} = 1, g_{tp} = g_{tn} = 0$). Also, Theorem 7.2.4 is an analogue of Narasimhan's (2015) result for multi-class classification:
3. Although recall $@k$ is intuitively instance-wise metric, as it can be calculated for every instance and label's coordinate and then averaged over them, it does not belong to general instance-wise weighted utilities, as defined in (3.9). Thus, without further discussion, the sentences above and below (3.18) seem contradictory;
4. p. 51: Blanchard's identifiability condition is vaguely stated;
5. There is a discrepancy between definition of PLT given in Chapter 9 (labels assigned to nodes) and that given in the paper Jasinska et al (2016)

which is cited in this context (labels correspond to leaves only) without any comment;
6. (2.25): In my opinion there is an important assumption here which is not explicitly stated, namely that $\check{y}_j \perp y_{-j}|x, y_j$ ($\check{y}_j$ is conditionally independent from all labels but y_j given x and y_j). In particular, this is assumed in Theorem 4.1.2.

Typos, small omissions and notational problems

1. G should be defined as $G = [g_1^T, g_2^T, \ldots, g_m^T]^T$ ($(m \times 4)$ matrix) and not $G = [g_1, \ldots, g_m]$ $(1 \times (4m))$ vector;
2. Definition 7.1.1: condition $\xi_1, \xi_2 \geqslant 0$ is missing;
3. Typos: p. 29: P is redundant in the definition of ILIR; (5.2) indices mixup: j should i in the first line; $j = 1$ should replace $i = j$ in the second and the third line; p. 11: prone to the nose (!); p. 95: erroneous definition of sigmoid function; p. 92: 'at the risk of suffering small regret': what is meant here, I believe, is the risk of increased regret; p. 68: $\mathcal{O}(1/\sqrt{n})$, not $\mathcal{O}(1/\sqrt{m})$; (7.5): the expectation is taken also wrt the randomisation pertaining to randomised classifier;
4. Notation: $\phi^j cov$ (p. 74) is very misleading: it suggests either multiplication by ψ^j or transformation by ψ^j of cov: it is neither of that;
5. The bound in (7.11) is rather awkwardly stated: if the two first summands of the bound are given as $\mathcal{O}$ terms, I do not see the point of giving explicit constant for the third;
6. Polish summary: nomenclature: 'regret' is equivalently called in English 'excess of the risk' which is much more revealing. In my opinion, it is preferable to call the concept in Polish as 'nadwyżka ryzyka' instead of used 'żal', which is emotionally-charged and not very evocative;

Specific comments and listed glitches (the number of which is surprisingly small given the length of the thesis) does not change my overall appraisal of the dissertation which is unequivocally positive. The thesis attests to Candidate's in depth-knowledge of Machine Learning and Optimisation Methods as applied to currently intensively researched problem of XMLC.

Conclusion

The thesis shows that the Candidate mastered proficiency of XMLC problem both on theoretical and computational level. Moreover, the novel results obtained by him contribute to understanding of its nature and to the solution of it. His sustained efforts during 7 plus years towards preparation of this thesis strongly indicate he is fully capable of conducting (fruitful) scientific work.

In view of the appraisal above, I attest that in my opinion Ph.D. thesis of Marek Wydmuch without any doubt fulfils all formal requirements for Ph.D. theses in the discipline of Information and Communication Technology (as stated in Law of Higher Education and Science, 2018 with amendments) and I will vote in favour of him defending it.

Moreover, in view of the significance of the results obtained discussed above, I recommend it for distinction.

Jan Michniczuk