View Meta-Reviews

Paper ID

838

Paper Title

Extreme Multi-label Regression and Ranking

META-REVIEWER #1

META-REVIEW QUESTIONS

1. Please recommend a decision for this submission.

Reject

3. Please provide a meta-review for this submission. Your meta-review should explain your decision to the authors. Your comments should augment the reviews, and explain how the reviews, author response, and discussion were used to arrive at your decision. Dismissing or ignoring a review is not acceptable unless you have a good reason for doing so. If you want to make a decision that is not clearly supported by the reviews, perhaps because the reviewers did not come to a consensus, please justify your decision appropriately, including, but not limited to, reading the submission in depth and writing a detailed meta-review that explains your decision.

The paper considers the problems of extreme multi-label regression and ranking and proposes an extension of Parabel, a recently proposed algorithm for these tasks (it also considers a new evaluation metric). The initial review scores were 6,6,4,4. The authors provided confidential comments to the AC indicating they were unhappy with the last two reviews. Following this, there was significant discussion among all the reviewers; overall, as indicated in the final reviews, it was felt that the contributions of the paper beyond prior work are not sufficiently significant for NeurIPS.

We thank all the reviewers for your comprehensive and helpful comments. We will incorporate your suggestions to improve the clarity, readability of the paper and make the paper error-free and self-contained. To reiterate the significance of our paper, please note that 1) XRR achieves SOTA performance on PSP@k metric which is an important accuracy metric in extreme classification literature (Table 4), 2) XRR has been successfully deployed for Dynamic Search Advertising on a prominent search engine and has shown impressive gains (Table 2), 3) paper introduces a novel extreme regression paradigm and metric, a novel labelwise inference paradigm and algorithm, and a novel theoretical analysis leading to generalization bounds with log dependence in the number of labels.

Update of Eq (1): We thank the reviewers for pointing out this mistake. The updated equation is as follows: XRMSE@$k = \sqrt{\frac{1}{k}\sum_{l=1}^{k}(y_{\hat{r}_l} - \hat{y}_{\hat{r}_l})^2}$.

Update of Eq (3): We are sorry for missing the label subscripts in Eq (3). The updated equation is as follows: $\min_{\mathbf{w}_n} \|\mathbf{w}_n\|^2 + \frac{C}{|\mathcal{I}_n|}\sum_{i\in\mathcal{I}_n} s_{ni}z_{ni}\log(1+\exp(-\mathbf{w}_n^\top x_i)) + s_{ni}(1-z_{ni})\log(1+\exp(+\mathbf{w}_n^\top x_i))$ where $z_{ni} \in \mathbb{R}$ is the maximum of the marginal relevances of all the labels in the subtree rooted at node n with respect to the ith point, $s_{ni}z_{ni}$ denotes the probability that ith point visits the node n and $s_{ni}(1-z_{ni})$ is the probability that ith point visits node n's parent but does not traverse to node n.

R1: **Optimality of XRR for XRMSE@k**: XRR optimizes an upper bound of XRMSE@k through weighted logistic regression as shown in Thm 4.2 and the proof in Thm 9.3 (line 599 in the appendix). **Contribution (3)**: XRR's pointwise inference is same as Parabel, but the labelwise inference is novel and can be found in Alg 1 of the appendix. **XRR Inputs & Outputs**: XRR training takes as input a set of training points of the form $\{\mathbf{x}, \mathbf{y}\}$ which obey the domain constraints noted in line 164 of our paper. Note that $\mathbf{y}$ is a real-valued vector and not binary-valued. The inputs for both pointwise and labelwise XRR prediction algorithms are the same: XRR model and a batch of test points. The pointwise prediction outputs k most relevant labels along with their relevance estimates for each test point, whereas the labelwise prediction outputs k most relevant test points along with their relevance estimates for each label. **Lines 143, 169 & 177**: Apologies for not being clearer. We intended to say "not all optimal rankings lead to good regression performance". Even a naive prediction algorithm which randomly picks k labels and assigns 0 for their relevance estimates can minimize XRMSE@k error with high probability since most labels have 0 relevance for a test point. Hence, minimizing XRMSE@k metric, on its own, is meaningless, but needs to be always coupled with a ranking metric. **Line 189**: XRR tree construction is the same as that of Parabel's and will be included in the Appendix. **Eq 3**: As noted in lines 201-210 of the paper and Eq (3) of this rebuttal, the latent z variables are instantiated from y_l variables prior to training. Eq 3 in each node models the probability that a test point travels to the node given that the point has already visited its parent. By multiplying all such node probabilities along a tree path, we get the marginal relevance for the label in which the path terminates. We will fix the other issues pointed out by R1 and are thankful for the same.

R2: **Parabel**: A detailed description of Parabel algorithm is provided in the appendix (Section 8). Lines 180-183 describe the extensions to Parabel that are result in the XRR algorithm. **Scalability of pairwise ranking methods**: Most large-scale, production-grade, recommendation systems are multi-staged (*e.g.* Rosset *et.al*, SIGIR'18). For example, in DSA, the L0 ranker uses simple word matching algorithms or highly scalable extreme classifiers to shortlist a small set (around 100) of most relevant ads for a query from millions of ads; while the further stages rerank these shortlisted ads more accurately by using more accurate but expensive algorithms. XRR is primarily targeted at L0 stage while IR-GAN and TF learning to rank suite are used in L1/L2 reranking stages, as described in these respective papers. The latter algorithms won't scale to millions of query, ad choices. **Batch labelwise inference**: XRR labelwise inference in DSA is run offline every 3-4 days on all the millions of ads in the DSA's ad corpus and outputs an inverse index of 5-10 most relevant ads for each of the query in XRR model. When a user asks a query on a search engine, it is looked up in the inverse index to get the list of relevant ads. Due to its offline nature, the labelwise inference allows batchwise prediction leading to much more efficient implementation possibilities. These details will be added in our next revision. **Use of tail classifiers**: Tail classifiers boost the PSP@k numbers by re-ranking the predicted tail labels. Please refer to section 6.2 of PfastreXML [22] for more details. **Use of 3 trees**: In our experiments, 3 XRR trees were found to give the best trade-off in terms of accuracy and efficiency: 1 tree is around 2% worse in WP@k and 5 trees are 1.7x slower to train and predict without any significant gains in WP@k. Similar trends hold for Parabel too. These results will be included in next revision. **WikiLSHTC reference**: This will be fixed. **MovieLens-138K reference**: The dataset was inspired by XLR [9] and created from https://grouplens.org/datasets/movielens/20m/ with a minor change in featurization. We will make this dataset public after the acceptance. None of the existing works report the results with the metrics discussed. We request R2 to increase the score and facilitate for the acceptance.

R3: We have fixed the mistake in Eq (1) above and will improve the clarity of writing in the next iteration. We will also include Parabel and metrics (WP@k & AUC@k) in the main article as mentioned above. We will restructure our tables for ease of reading. Lastly, thanks for pointing out the issue in the References section, it was a mishap and will be fixed. We request R3 to kindly reconsider the decision based on the merits of our paper listed at the top.

R4: We have fixed the mistakes in Eq (1) & (3) above. We will improve the writing as mentioned at the start of the rebuttal. **Confusion in the experiments & metrics**: As observed in lines 313-320 of paper and lines 23-26 of this rebuttal, the XRMSE@k metric will be a fair comparison only when coupled with a ranking metric like WP@k. Lines 78-89 in paper address your concern about the value of XRMSE@k. We request R4 to kindly reconsider the decision based on the merits of our paper listed at the top.

View Reviews

Paper ID

838

Paper Title

Extreme Multi-label Regression and Ranking

Reviewer #1

Questions

1. Contributions: Please list three things this paper contributes (e.g., theoretical, methodological, algorithmic, empirical contributions; bridging fields; or providing an important critical analysis). For each contribution, briefly state the level of significance (i.e., how much impact will this work have on researchers and practitioners in the future?). If you cannot think of three things, please explain why. Not all good papers will have three contributions.

The paper proposes XRR, an extension of Parabel to extreme regression and ranking. The basic idea is to compute a bound on the divergence between the true and predicted marginal label probabilities, and use this as objective to optimise during training. It also proposes XRMSE, a metric for extreme regression problems. A detailed experimental comparison shows improvements owing to the XRR method.

2. Detailed comments: Please provide a thorough review of the submission, including its originality, quality, clarity, and significance. Hover over the "?" next to this prompt to see a brief description of these metrics.

Update:

Thanks for the response. It seems there is generally agreement (which the authors accept) that the paper's presentation could be significantly improved.

The response cleared up one question I had, namely, the connection between the two contributions of the paper. I am less sure about the nature of the input to their algorithm -- I suspect it must be an instance and its historical clickthrough rate, but the response mentions that y respects the preceding domain constraints that allow it to be any real number, which is confusing.

Overall, a fairly major revamp seems necessary to make the algorithm clear. I do think that the real-world results in Table 2 are commendable, and so encourage the authors to work on the presentation of ideas.

=================

The paper proposes XRR, an extension of Parabel to extreme regression and ranking. The basic idea is to compute a bound on the divergence between the true and predicted marginal label probabilities, and use this as objective to optimise during training. It also proposes XRMSE, a metric for extreme regression problems. A detailed experimental comparison shows improvements owing to the XRR method.

The paper is well-motivated: extreme regression and ranking are important problems, and have received less attention than their classification counterpart. The need to exploit and estimate relevance weights for online advertising, which is presented as a key motivation of the present work, is a convincing real-world application of the techniques developed. The paper also conducts a fairly thorough comparison of the proposed method against a number of baselines.

In terms of technical contributions, the extreme RMSE metric is essentially the RMSE restricted to the top-k labels. Though not particularly profound, this is intuitive enough, and does not appear to have been explicitly put forth in the literature. The XRR model is a reasonable extension of Parabel. The core idea is to optimise a bound on the KL divergence between the true and predicted marginal label probabilities. This bound is expressed in terms of the marginal probabilities of the intermediate nodes, and is a simple consequence of the properties of the KL divergence. One can then directly optimise this bound, as it is expressible as a function of individual nodes in the internal tree maintained by Parabel. This is again not particularly profound, but is a reasonable means of addressing an important real-world problem.

There were a few points of confusion upon a first read, some of which remain:
- there is an apparent disconnect between the two contributions. In particular, it does not seem that XRR uses the XRMSE metric directly during optimisation. Instead, it uses a weighted logistic regression, where the weights are provided as input (e.g., historical CTRs), and the logistic probabilities are treated as estimates of the weights (e.g., out-of-sample CTRs). From my understanding, the XRMSE is used to measure performance of estimated relevance weights. It was not clear how (exactly) optimisation of this quantity is encouraged in the objective used for training.

- the precise set of inputs to and outputs from the XRR algorithm are unclear, especially to someone less familiar with Parabel. Certainly (3) suggests that one has access to instances and their labels, but the preceding discussion is suggestive that one also has access to some form of real-valued relevance score. From my understanding, it is assumed that one has access to (noisy estimates of) the marginal label probabilities, y_l. I did not follow the discussion of how one converts this into the scores z that are used in (3). I further did not follow how precisely (3), which is presumably solved for each node, is used to arrive at the final predictions.

Regarding model outputs, it is worth explicating that (in my understanding) the goal is to learn model estimates of the marginal label probabilities, with the aim of smoothing existing estimates and providing generalisation to unseen samples.

- something that confused me initially was the claim that XRR is a regression method, as opposed to a classification method that relies on thresholding real-valued scores (as in vanilla logistic regression). Upon re-reading, the resolution (in my understanding) is that the authors view the problem of learning from samples (x, y) for binary y with the goal of estimating P(y = 1 | x) as a regression problem. Class-probability estimation is sometimes used to refer to the problem where one receives binary targets, but wishes to predict the probability P(y = 1 | x) e.g., by vanilla logistic regression. Regression is sometimes used to refer to the case where one has access to real-valued (not binary) targets.

As a general comment, the writing is generally good, but the presentation has some scope for improvement. For example, the Introduction discusses several related topics in succession, delineated by inline headings. There is not much in the way of supporting text to present the big picture, or connect these themes; this leads to some choppiness in flow. For example, the Objective subsection makes mention of pointwise and labelwise inference, but these terms are only defined after two further subsections. The first two paras of Section 3 also seemed a little repetitive, as by this stage the DSA and extreme regression problems have been discussed in both Secs 1 and 2.

Minor comments:
- in the definition of xRMSE, it does not appear that the ranks of the predicted targets are used.
- there are some grammatical issues:
* "generalises the extreme multi-label learning to" -> delete "the"
* "could be available" -> "may be available"
* "modifies Parabel algorithm" -> "modifies the Parabel algorithm"
* "making it the good choice to" -> "making it a good choice to"
* "even after a heaving down-sampling" -> "heavy"?

- contribution (3), about efficient inference schemes for XRR, seem to rely on the same inference scheme for Parabel in the labelwise case?
- Line 143, "optimal ranking doesn't mean good regression" -> true, but a little confusing, since there is no unique optimal ranker but a set of them, unique upto strictly monotone transformation. I think there is an optimal ranker which has optimal regression performance, i.e., predicting E[y_l | x]?
- Line 166, Pi (set of permutations) is undefined.
- Line 169, "Minimizing XRMSE@k metric...results in the best performance" -> not clear what the definition of "best performance" is.
- Line 177, "trivial solutions such as ranking k irrelevant labels..." -> not clear what is meant here.
- it might make more sense to first present the XRR algorithm, which I gather is intended to be the primary contribution of the work.
- Line 189, some background on the clustering procedure used to construct the internal tree, even if in the Appendix, would be useful.
- Theorem 4.1, it is prudent to have a sentence or two on the "unvisited node assumption".
- Equation 3, it is not explicated that the logistic loss is used by virtue of its excess risk corresponding to a KL divergence.
- Equation 3, explicate that you are modelling \hat{z}_{lh} using sigmoid(w^T x).
- Theorem 4.2, one must need to assume that the linear model used in (3) is well-specified for the underlying marginal label probabilities, else the excess risk cannot go to zero?

3. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold. I tend to vote for accepting this submission, but rejecting it would not be that bad.

4. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

5. Improvements: What would the authors have to do for you to increase your score?
Clarify relationship between the contributions, and details of how one computes final model predictions for XRR.

Reviewer #2

Questions

1. Contributions: Please list three things this paper contributes (e.g., theoretical, methodological, algorithmic, empirical contributions; bridging fields; or providing an important critical analysis). For each contribution, briefly state the level of significance (i.e., how much impact will this work have on researchers and practitioners in the future?). If you cannot think of three things, please explain why. Not all good papers will have three contributions.
The paper contributes a new algorithm for classification and recommendation tasks in large-scale settings. Provides empirical comparisons in a diverse set of tasks and datasets with quite good analysis and discussion.

The algorithm is an extension of a previous one but seems to perform quite good in the different task. Results are in the SOTA for some of the datasets.

2. Detailed comments: Please provide a thorough review of the submission, including its originality, quality, clarity, and significance. Hover over the "?" next to this prompt to see a brief description of these metrics.
The authors present a new algorithm for large-scale classification and recommendation. It is based on a previous algo (Parabel) which extends to accommodate for real values in the prediction task (e.g.

relevance feedback). The authors compare with several other approached in multiple datasets.

The paper even though dense it reads quite well. One issue is the constant reference to the Parabel algo that is extended in this work. This is quite annoying as the paper is not self-contained and one has to consult an extra paper. I would suggest the authors to give a brief description of Parabel. Also, it needs some clarification of the exact extensions over the Parabel model.

The authors mention in their related work that for paiwise ranking techniques do not scale. This statement does not take into account the latest developments in neural models like IR-GAN which can scale to very large problems or the most recent TF learning to rank suite. Authors seems to totally ignore such approaches.

Why labelwise inference has to happen for a bunch of test point? Is this necessary? How is this implemented in the real-world scenario you describe?

What exactly is the use of the tail classifiers to boost performance? Why this information is not included in the paper?

Why for the proposed approach you use 3 different trees? Is the same happening for Parabel? How you optimized this number? Do you have results with just one tree?

The reference to the WikiLSHTC dataset is not correct. Please, include the original reference from the LSHTC challenge.

Can you please provide the exact link of the MovieLens dataset you used?

Performance on the dataset seem good with respect to other approaches. Although, for the recommendation tasks the authors could have used more powerful approaches than RankSVM. Are there any other reported results on the MovieLens dataset from other researchers?

Overall a good paper.

========================
Update after author responses.

I would like to thank the authors for the detailed replies. I acknowledge the adequate presentation of Parabel. Which also raises a question. What if no tree is given? What about DAG structures? Many open questions here which are not clarified.

In terms of presentation I would advise the authors to include a concise presentation of Parabel. Also, adding stronger baselines in the recommendation datasets would be helpful to evaluate the model.

3. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold. I tend to vote for accepting this submission, but rejecting it would not be that bad.

4. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

5. Improvements: What would the authors have to do for you to increase your score?
Improve the presentation and give more details on the approaches. Also, if possible add comparisons for the recommendation tasks with other approaches.

Reviewer #3

Questions

1. Contributions: Please list three things this paper contributes (e.g., theoretical, methodological, algorithmic, empirical contributions; bridging fields; or providing an important critical analysis). For each contribution, briefly state the level of significance (i.e., how much impact will this work have on researchers and practitioners in the future?). If you cannot think of three things, please explain why. Not all good papers will have three contributions.

1 This paper proposes a novel metric XRMSE for measuring extreme regression performance.
2 This paper proposes a novel extreme multi-label algorithm named XRR.
3 Many comparison experiments on large scale data sets have been implemented.

2. Detailed comments: Please provide a thorough review of the submission, including its originality, quality, clarity, and significance. Hover over the "?" next to this prompt to see a brief description of these metrics.

This paper proposes an extreme root mean squared error as a novel metric for extreme multi-label learning. The idea is very simple and trivial, it is just the top k labels' root mean squared error. In addition, $\hat{r}_l$ has not been concerned in Eq. (1), why did you define them? Have you missed something in the Eq. (1)?

The written of this paper is not clear. This paper proposes a novel XRR model based on Parable algorithm. Since Parable algorithm is very important, it is better to include it in the paper not just throw it into the supplementary. For WP@k, it is also an important concept that should be included in the paper. In the part of contributions, you said 3 contributions but numbered from (1) to (4).

The organization of this paper is not good for me. For example, Table 1 is so tiny. I know that the page limit is not allowed to show all your things in 8 pages, but I guess you can find a better way to reorganize them.

Last, the references of this paper missed a lot of things. For example, do [3, 6, 8, 9 ...] publish in a journal or a conference? Where are their volume (issue) or pages information?

3. Please provide an "overall score" for this submission.

5: Marginally below the acceptance threshold. I tend to vote for rejecting this submission, but accepting it would not be that bad.

4. Please provide a "confidence score" for your assessment of this submission.

2: You are willing to defend your assessment, but it is quite likely that you did not understand central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

5. Improvements: What would the authors have to do for you to increase your score?

Is there any error in Eq. (1)?

Reviewer #4

Questions

1. Contributions: Please list three things this paper contributes (e.g., theoretical, methodological, algorithmic, empirical contributions; bridging fields; or providing an important critical analysis). For each contribution, briefly state the level of significance (i.e., how much impact will this work

have on researchers and practitioners in the future?). If you cannot think of three things, please explain why. Not all good papers will have three contributions.

In this paper the authors propose a new error measure and algorithm for extreme regression

2. Detailed comments: Please provide a thorough review of the submission, including its originality, quality, clarity, and significance. Hover over the "?" next to this prompt to see a brief description of these metrics.

In this paper the authors propose a new metric and a new algorithm for extreme regression, in which the goal is to regress from a given input many different outputs that then are ranked and only the largest are useful or shared.

The proposed loss function is mentioned in Section 3. The authors after presenting the loss function the mentioned that it cannot be trained in isolation, because the ranking also needs to be computed correctly. But the loss function does not depend on the estimated ranking $\hat{r}$, but the true ranking r, needing to optimize WP too is not due to the XRMSE loss but a problem of the subsequent algorithm. WP loss depends on $\hat{r}$ and not r, which explain why we need the mixed

The new algorithm is described in Section 4.1 and specifically on Equation 3. But in that equation, it is unclear the role of z_n and s_n. They seem to be scalars that are identical for all training examples. The sum is with respect to i and w_n are the parameters, so a fixed z_n for all samples does not make sense, also z_n seems to play the role of the label in the logistic regression solution in equation 3. Also a fixed s_n would not have any effect in the optimization. Finally neither z_n or s_n are defined in the paper. My take is that they should depend on i too. This presentation of the algorithm makes it very complicated to understand what the authors are actually proposing and why it is any good.

The experiments are also confusing. First the proposed metric never favors the proposed algorithm, except in one case in Table 5. These sounds weird, the authors algorithm fares much better with the other loss function defined in the supplementary material. Why do they propose it and what it is its value?

Also, the variability between the different error measures is huge. I would expect some consistency between them. For example, in Table 3 for database WikiLSHTC-325K-p, the proposed algorithm is better for the WP@5 and AUC@5 error measures, but much worse for the proposed error measure. Actually, the algorithm that beats all of the others in this case by a significant margin, it is the worse in the other two measures. This is an example, but in many other cases it happens too. For me this is puzzling.

My take is that the paper needs mayor rewriting before it can be accepted for publication.

After Feedback: Thank you for pointing out the changes in the equations, but there are still quantities not well-defined. I think the paper needs to be rewritten in a way that the true merits of the paper stand on its own. Right now it is very complicated to follow.

3. Please provide an "overall score" for this submission.

4: An okay submission, but not good enough; a reject. I vote for rejecting this submission, although I would not be upset if it were accepted.

4. Please provide a "confidence score" for your assessment of this submission.

3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

5. Improvements: What would the authors have to do for you to increase your score?

I think the paper needs to be written better.