December 9, 2020

INESC TEC
Campus da Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto
Portugal

Article id: 53280
Article Title: Ordinal Losses for Classification of Cervical Cancer Risk

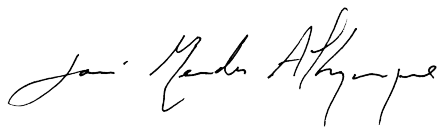
Dear Editors,

We thank the reviewers for their generous comments on the manuscript and we have edited the manuscript to address their concerns.

All of the code we wrote is available and I have included the link throughout the paper to the appropriate code repository.

We look forward to hear what your thoughts are about the updated manuscript.

Yours sincerely,



Tomé Mendes Albuquerque

On behalf of all authors.

Editor:

Regarding Reviewer 3's comments, given the focus on cervical cancer imaging, it may not be possible or necessary to extend the collection of datasets used in the study. However, significance tests need to be performed on the estimates of predictive performance, and measures of variance need to be provided. This should be feasible even with deep learning given the small size of the dataset.
The exact experimental protocol used in the study remains unclear to me. It is important to state exactly how much data is used for training, validation (i.e., parameter tuning and early stopping), and testing. If stratified k-fold cross-validation is performed to establish the final performance estimates, then parameter tuning needs to be performed separately for each of the k runs, making sure that information from the test set of run k does not influence hyperparameter choice for run k in any way.

We agree with your concern and we implemented stratified 10-fold cross-validation and we also implemented nested k-fold to do parameter tuning (λ of our proposal loss). We updated the manuscript including in "Train" subsection more information about cross-validation methodologies used during the train.

In my opinion, given the current results, where OE appears very competitive with the proposed new ordinal loss functions, the paper should deemphasise the novel loss functions and instead focus on the possibility that using ordinal methods improves results on this cancer data (assuming superiority holds after significance testing).
An empirical comparison of different deep ordinal classification approaches (including the new ones) on this data seems a valuable contribution. In this regard, the suggestions by Reviewer 2 need to be addressed, particularly the simple baseline using the "expectation trick" and the published deep ordinal methods cited in the review.

We agree with your concern that we should deemphasise the novel loss functions and focus on the idea that using ordinal methods improves results on this cancer data. We updated the manuscript by changing article title to "Ordinal Losses for Classification of Cervical Cancer Risk" and by adding more details between parametric and no-parametric losses. We also write in "results" and "conclusion" sections a critical analysis of the results obtained by nominal losses (CE) VS ordinal losses and we also analyze the differences between parametric and non-parametric losses in relation to their performance.

Reviewer 1: Pingjun Chen

Basic reporting: The writing is unambiguous and easy to follow. Background and related work are clear and rather detailed. "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss" is suggested to add to the related work.
Experimental design: Considering the ordinal nature of the pap smear cell classification, the authors propose a non-parametric ordinal loss to promote the output probabilities to follow a unimodal distribution.
Validity of the findings: The authors experiment with the proposed loss on the Herlev dataset on multiple CNN architectures. In addition, the authors compare several other losses. Experiments show the effectiveness of the proposed methods.
Comments for the author: The authors propose a novel ordinal loss for the pap smear cell classification. They focus on promoting the unimodal distribution of the output probabilities, which is a good insight into the ordinal classification problem. The experiments and evaluations well demonstrate the idea.

Reviewer 2: Christopher Beckham

To elaborate on my 'stronger baselines' point, it seems like the main reason why this loss was proposed is because we do not necessarily want a distribution that is purely unimodal (like in the case of PU). Perhaps that is partly because (1) the conditional probability distribution should not be modelled by a unimodal distribution; and/or (2) using PU (i.e. a binomial distribution) would be too constraining since the variance cannot be easily controlled.

To address point (2): a variance-controlled version of the binomial distribution does exist – called the Conway-Maxwell Binomial (CMB) [2,3] – which has a variance-controlling term. That means that your network could be modified to have two outputs: (p, v) = f(x), and then you can maximise the log-likelihood of a CMB distribution. (A more heuristic version of this was proposed in [1], but it's essentially CMB.)
Secondly, to address point (1): why not just infer a mixture distribution between a regular softmax (CE) distribution and a unimodal (PU) one? For instance, suppose your network was modified to have two outputs: $p\_s(y|x) and p\_u(y|x)$, where p_s denotes a regular softmax distribution and p_u the unimodal one, you could simply (for some alpha in [0,1]) construct a mixture distribution between the two: $p(y|x) = alpha * p\_u(y|x) + (1 - alpha) * p\_s(y|x)$. alpha could either be a hyperparameter to tune, or you might even be able to get away with making it a learnable parameter as part of the network. This would make for a somewhat interesting method, since a high value of alpha would put more weight on p_u, essentially acting as a strong regulariser on the distribution.
Thirdly, the more competitive version of the simplest baseline (CE) would be to do a post-hoc label prediction based on the 'expectation trick' found in [1]. Essentially, for some $p(y|x)$, if we assign to each class an integer label $[1, 2, \ldots, K]$, we take the expected value of this distribution by computing the dot product between $p(y|x)$ and $[1, 2, \ldots, K]$, and round the result to the nearest whole integer. This basically uses all of the probability mass in $p(y|x)$ to make a prediction.

In summary, I would compare your proposed technique to: - A more competitive CE using the expectation trick

We agree with your concern and we implemented the expectation trick for all losses and architectures. We updated the manuscript by adding two new tables (Table A3. and A4. with the aggregate results for **4 and 7 class** problem, averaged for 10 folds.

- Use a parametric unimodal (PU) method using the CMB distribution

We found your proposal very interesting and decided to implement the new parametric loss using a Poisson distribution based on your article: "Unimodal Probability Distributions for Deep Ordinal Classification". We named this loss as Poisson Unimodal (PU). We updated the manuscript by adding information about this loss in "Related Work" section and also by adding in all tables the results for this loss.

- Experiment with using a mixture distribution between p_u and p_s

We found your comment very interesting for future works, however we decided to not implement in this article because we thought we would deviate from the central focus of the article.

Other less significant points: - While it was appreciated that the authors tried out a vast range of architectures, perhaps it would make for a better presentation if the number of learnable parameters was stated for each of these architectures. You could then explore performance vs # parameters. It seems like the dataset you have used is extremely tiny, and having excessively large networks could degrade generalisation performance here. If it saves you computational resources, I don't think some of these architectures are strictly needed in the analysis: for instance AlexNet and VGG, which were superceded by ResNets (and for good reason). - It would be interesting to explore the case where you don't start off with pre-trained ImageNet weights. While I would expect such a network to very easily overfit, it can be controlled with sufficient regularisation (weight decay), and also allow you to explore the effect of having a severely constrained distribution (i.e. PU) in a 'low data' setting.

We agree with your concern, we also want to explore in future works the performance vs # parameters among the different architectures. We run our models across a large number of architectures to prove the robustness of our proposal loss regardless the architecture.

Reviewer 3:

Experimental design Only averages of 5 folds are given, in order to show the robustness I suggest to provide results of multiple experiments (e.g. 10), then averages and variance / standard deviation or box plots. And in addition I suggest to perform statistical significance tests on the prosed and tested algorithms.
Validity of the findings The proposed algorithm cost functions are straightforward. It would be a surprise if ordinal classification can benefit from these cost functions in general. Only a rigorous statistical evaluation of the proposed cost functions based on 10 or more data sets utilising more complex statistical evaluation (e.g. Wilcoxons test) in comparison with the other approaches could prove the strength of the proposed algorithm.

We agree with your concern and we implemented stratified 10-fold cross-validation and trained again all the models. We also updated the manuscript Tables of results. Furthermore, the results are compared to the best loss result and a statistical test is used with the hypothesis of them being the same or not. A p-value of 0.1 is used with a one-sided paired t-test.