Labelling Strategies for Hierarchical Multi-Label Classification Techniques

Isaac Triguero ISAAC.TRIGUERO@IRC.VIB-UGENT.BE
Ghent University, Department of Internal Medicine, Ghent, Belgium
VIB Inflammation Research Center, Data Mining and Modelling for Biomedicine group, Zwijnaarde, Belgium

Celine Vens CELINE.VENS@KULEUVEN-KULAK.BE
KU Leuven Kulak, Department of Public Health and Primary Care, Kortrijk, Belgium

Keywords: hierarchical multi-label classification, threshold optimisation, hierarchical loss, HMC-loss, F-measure

Abstract

Many hierarchical multi-label classification systems predict a real valued confidence score for every label, rather than an actual label set. In this study, we present and compare different strategies to perform automatic labelling.

1. Introduction

Hierarchical multi-label classification (HMC) involves predicting multiple and partial paths in a hierarchy of labels. Examples are found in biology (e.g., gene function prediction (Clare & King, 2001)), text or image classification (Boutell et al., 2004), etc. Several HMC algorithms have been proposed in the literature, see for example (Cerri et al., 2015). They exploit the label set hierarchy when labelling instances. These systems also ensure (implicitly or using post-processing) that the hierarchy constraint is fulfilled in the predictions they make: whenever a class is predicted, its parent and ancestor classes are also predicted.

Rather than predicting an actual label set, most of the HMC algorithms actually predict a real valued prediction score p_i for every label l_i, that reflects the confidence that an instance should be annotated with label l_i. These values can be easily converted into a label set by applying a threshold on them: if p_i is above some threshold t_i, then the instance is predicted to belong to class l_i, otherwise not. To ensure that the predictions fulfil the hierarchy constraint, it suffices to choose $t_i \leq t_j$ whenever l_i is a super class of l_j.

Often, the decision as to which thresholds to choose is left to the end user, and the predictive performance of the classification algorithms is evaluated in a threshold independent way, for example by using precision-recall curves. However, in some situations, it is preferable or necessary to fix the thresholds. For instance, the gene function prediction task may be part of a larger pipeline of experiments, or the predicted image labels may be used as tags in image retrieval systems to locate images of interest.

In a recently published article (Triguero & Vens, 2016), we present and evaluate different alternatives to perform the actual labelling in hierarchical multi-label classification. HMC studies that fix the thresholds typically choose one threshold shared by all labels. In the non-hierarchical multi-label setting, however, studies exist that choose a separate threshold per label (Pillai et al., 2013). In our article, we investigated how these two options compare in HMC.

2. Method

Non-hierarchical optimisation techniques can not be straightforwardly applied in the HMC context, because of the aforementioned hierarchy constraint, and thus, we proposed adapted techniques. Depending on the context, the user may want to set the thresholds such that the resulting classifier maximises predictive performance or such that training set properties (such as class distribution or label cardinality) are reflected in the predictions. We proposed threshold selection algorithms for both strategies, both under a single and

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

a multiple threshold selection scheme. We used three performance measures used in HMC to compare a predicted label set to a true label set: H-loss, HMC-loss and micro-averaged F_1-measure.

3. Results

We performed an empirical comparison on 10 HMC datasets. The number of labels ranges from 56 (enron) to 4125 (cellcycle_GO). Fig. 1 shows a comparison of five proposed labelling strategies evaluated using micro-averaged F_1. The average ranks diagram shows the best performing approaches to the right; all approaches connected by a red line are not statistically significantly different from each other. We see that there is no significant difference between the four best methods. Fig. 2 shows a run time comparison for the two best performing methods, which select the thresholds such that the F_1 measure over a validation set is optimised. Surprisingly, the multi-threshold selection approach, which selects a different threshold per label, is much faster. This is because the hierarchy constraint drastically decreases the search space.

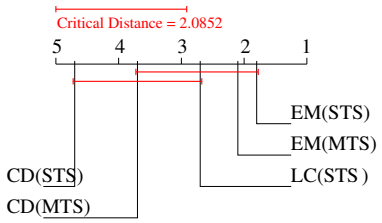


Figure 1. Average ranks diagram for micro-averaged F_1 comparing single (STS) and multiple threshold selection (MTS) for optimising the evaluation measure (EM), simulating class distributions (CD) and simulating label cardinalities (LC).

As a side contribution, we have also shown that the widely used H-loss measure (Cesa-Bianchi et al., 2006) has a number of undesirable properties. For example, it fulfils none of the requirements for hierarchical evaluation measures listed by Kiritchenko et al. (2006), although it was designed specifically for HMC. Moreover, it favours empty predictions.

4. Conclusion

We conclude that, although selecting different thresholds per class has not been considered before in HMC, it provides a valid alternative, always resulting in better or equally good label sets than the single-threshold variant. Moreover, it always results in smaller computation time, because of the hierarchy constraint.

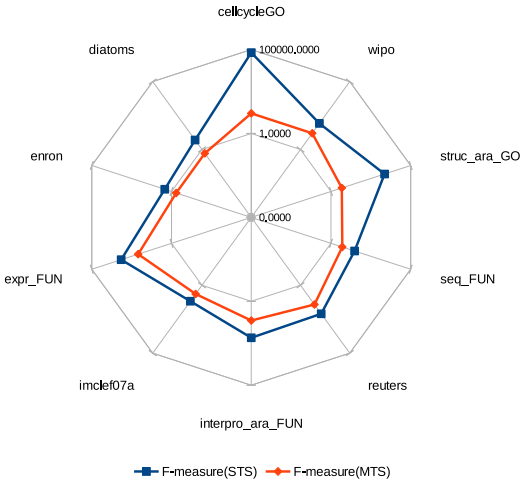


Figure 2. Run time comparison for EM(STS) and EM(MTS) from Fig. 1. Note the logarithmic scale.

References

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, *37*, 1757–1771.

Cerri, R., Pappa, G. L., Carvalho, A. C. P., & Freitas, A. A. (2015). An Extensive Evaluation of Decision Tree-Based Hierarchical Multilabel Classification Methods and Performance Measures. *Computational Intelligence*, *31*, 1–46.

Cesa-Bianchi, N., Gentile, C., & Zaniboni, L. (2006). Incremental algorithms for hierarchical classification. *J. Mach. Learn. Res.*, *7*, 31–54.

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. *Proc. of the 5th ECML/PKDD* (pp. 42–53). London, UK: Springer-Verlag.

Kiritchenko, S., Matwin, S., Nock, R., & Famili, A. (2006). Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in artificial intelligence*, vol. 4013 of *LNCS*, 395–406. Springer Berlin Heidelberg.

Pillai, I., Fumera, G., & Roli, F. (2013). Threshold optimisation for multi-label classifiers. *Pattern Recognition*, *46*, 2055 – 2065.

Triguero, I., & Vens, C. (2016). Labelling strategies for hierarchical multi-label classification techniques. *Pattern Recognition*, *56*, 170 – 183.