Hierarchical Text Classification Based on Separation in the Data or Feature Space

Dehghani, M.; Azarbonyad, H.; Kamps, J.; Marx, M.

Publication date
2016

Document Version
Final published version

Hierarchical Text Classification Based on Separation in the Data or Feature Space

Mostafa Dehghani[1] Hosein Azarbonyad[2] Jaap Kamps[1] Maarten Marx[2]

[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2]Informatics Institute, University of Amsterdam, The Netherlands
{dehghani,h.azarbonyad,kamps,maartenmarx}@uva.nl

1. INTRODUCTION

Hierarchy is a common and effective way of organizing data and representing their relationships at different levels of abstraction. However, hierarchical data dependencies cause difficulties in the estimation of "separable" models that can distinguish between the entities in the hierarchy. Extracting separable models of hierarchical entities requires us to take their relative position into account and to consider the different types of dependencies in the hierarchy. In this paper, we present an investigation of the effect of separability in text-based entity classification and argue that in hierarchical classification, a separation property should be established between entities not only in the same layer, but also in different layers.

2. RESULTS

There are two types of dependencies in the hierarchies: i) *Horizontal dependency*, which refers to the relations of entities in the same layer. For example the dependency between siblings which have some commonalities in terms of being descendants of the same entity. ii) *Vertical dependency*, which addresses the relations between ancestors and descendants in the hierarchy. For example the relation between root and other entities. Due to the existence of two-dimensional dependencies between entities in the hierarchy, modeling them results in overlapping models that are do not clearly distinguish the different entities. As a concrete example, consider a simple hierarchy of a multi-party parliament—with MPs, parties, and status, that is, government or opposition—which determines different categories relevant to the different layers of membership in the parliament, and we can classify these entities based on text of their speeches in parliament.

Separation in the data or feature space is a favorable property that not only helps to improve for ranking or classification algorithms, but also brings out characteristic features for human inspection. Figure 1 illustrates two different ways of modeling two entities, in the status layer of the parliamentary hierarchy, i.e., government and opposition. Each model is a probability distribution over terms (language model) based on the speeches given by all the members in the corresponding status. In each figure, we sort the terms based on their weights in one of the models, and plot the other in the same order. Although distributions over terms in Figure 1 (left) for two classes are different, they do not suggest highly separable representations for classes. However, estimated language models in Figure 1 (right) provide highly separable distributions over terms for two classes, identifying the characteristic terms that uniquely represent each class, and can be directly interpreted. Moreover, the language models in Figure 1 (right) select a small set of characteristic features, making it easy to learn effective classifiers for classes of interest.

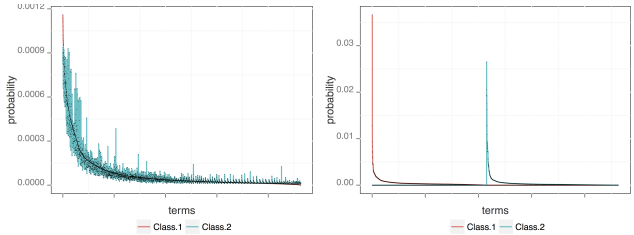

Figure 1: Probability distribution over terms for data in two different classes, sorted based on the term weights in one of the classes: a non-separable representation of data (left) and a well -separable representation of data (right).

3. CONCLUSIONS

Our main findings are the following. First, we analyse the importance of separability on the data representation in the task of classification and based on that, we introduce a "Strong Separation Principle" for optimizing expected effectiveness of classifiers decision based on separation property. Second, we present Hierarchical Significant Words Language Models (HSWLM) which capture all, and only, the essential features of hierarchical entities according to their relative position in the hierarchy resulting in horizontally and vertically separable models. Third, we validate our claims on real world data and demonstrate that how HSWLM improves the accuracy of classification and how it provides transferable models over time. Although discussions in this paper focus on the classification problem, the models are applicable to any information access tasks on data that has, or can be mapped to, a hierarchical structure.

REFERENCES

[1] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. On horizontal and vertical separation in hierarchical text classification. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 185–194, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4497-5. . URL http://doi.acm.org/10.1145/2970398.2970408.

*This is the abstract of Dehghani et al. [1], receiving the *Best Paper Award* at ACM ICTIR'16.

DIR'16 November 25, 2016, Delft, The Netherlands.