Ranked Recall: Efficient Classification by Learning Indices that Rank

Omid Madani
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504, USA
madani@yahoo-inc.com

Michael Connor*
Department of Computer Science
University of Illinois at Urbana-Champaign
N. Goodwin Ave, Urbana, IL 61801, USA
connor2@uiuc.edu

A fundamental activity of intelligence is to efficiently detect to which of myriad categories a given entity belongs. The problem occurs in many incarnations and applications, including: (1) categorizing web pages into the Yahoo! topic hierarchy (http://dir.yahoo.com) [MGKS07, LYW+05], (2) prediction problems [Mad06, Goo01, EZR00], and (3) determining the visual categories for image tagging and object recognition [WLW01, FP03]. Furthermore, ideally we desire systems that *efficiently learn to efficiently classify*. In particular, we would like to ensure that both learning of categories and categorization of items be efficient in their usage of time and space. However, these tasks present a number of challenges for learning:

- Large or practically unbounded training sets.
- Large dimensionalities (thousands and beyond).
- Large numbers of categories (thousands and beyond).

In this work, we explore an approach based on learning an index of features into the categories. An index is a sparse weighted bipartite graph that connects each feature to zero or more categories. During classification, given an instance, the index is looked up much like a typical inverted index for document retrieval would be: active features of the instance are used for the index look up, and categories are retrieved and ranked by the scores that they obtain during retrieval. We term this process *ranked recall* (of categories). The ranking and the category scores can then be used for category assignment.

We design our online algorithms to efficiently learn an index that accurately and efficiently ranks. We compare against one-versus-rest and top-down (hierarchical)

*Portion of this research performed while the author was at Yahoo! Research.

training and classification methods, using both perceptrons and support vector machines. In our experiments on a word prediction problem and six text categorization data sets, we find that the index is learned in seconds or minutes. Other methods can take orders of magnitude longer depending on the number of instances and classes. We achieve accuracies that are competitive and at times better than the best of the other methods. We gain significantly in terms of both space and time efficiency, during training as well as categorization times.

This research builds on the idea of index learning when the number of classes is huge [MGKS07]. However, in that work the objective was not accurate ranking, and the techniques relied on binary classifiers to achieve the best accuracies. We have observed, somewhat unexpectedly, that we no longer require classifiers.

References

[EZR00] Y. Even-Zohar and D. Roth. A classification approach to word prediction. In *Annual meeting of the North Amercian Association of Computational Linguistics (NAACL)*, 2000.

[FP03] D. A. Forsyth and J. Ponce. *Computer Vision*. Prentice Hall, 2003.

[Goo01] J. T. Goodman. A bit of progress in language modeling. *Computer Speech and Language*, 15(4):403–434, October 2001.

[LYW+05] T. Liu, Y. Yang, H. Wan, H. Zeng, Z. Chen, and W. Ma. Support vector machines classification with very large scale taxonomy. *SIGKDD Explorations*, 7, 2005.

[Mad06] O. Madani. Prediction games in infinitely rich worlds. In *Utility Based Data Mining Workshop (UBDM at KDD) and Y! Research technical report*, 2006.

[MGKS07] O. Madani, W. Greiner, D. Kempe, and M. Salavatipour. Recall systems: Efficient learning and use of category indices. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[WLW01] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.