Predicting Tags for Stack Overflow Questions

Sayan Pal, Ammar Shaker, and Eyke Hüllermeier

Intelligent Systems Group
Paderborn University

Stack Overflow (https://stackoverflow.com/) is one of the major community-driven Question and Answer (Q&A) websites, focusing on topics related to computer programming. It has nearly 7 million users, who ask more than 6700 questions every day. Each question can be associated with up to five different tags, which serve as metadata to facilitate information retrieval.

In this paper, we consider the problem of supporting this process through automatic tagging. From a machine learning point of view, we are facing a problem of Extreme Multi-Label Classification (XMLC), as Stack Overflow allows for choosing from several thousands of tags. Besides, instead of learning in a standard batch mode, it is desirable to learn incrementally on the continuous stream of questions entering the system, with the capability to capture changes and drifts in the data; for example, many tags (such as 'facebook') have a lifetime, first gaining popularity, then reaching a peak and eventually diminishing.

Thus, we end up with an extremely challenging problem of XMLC for data streams with a non-stationary set of labels. To tackle this problem, we build on an XMLC method based on probabilistic label trees (PLT), which has recently been proposed in [1]. We extend this approach in two directions. First, instead of specifying the entire PLT beforehand, we develop an adaptive version that starts with only a single node and expands the tree whenever a new label is observed in a training example. As a second contribution, we further improve adaptive PLTs through stream-based boosting [2]. More specifically, we apply the online boosting method by Oza and Russell, which we tailor for minimizing the F-measure as a performance metric (instead of 0/1 loss, which is not appropriate in the context of XMLC).

In addition to the methodological contributions, we present empirical results based on extensive experiments with real data from Stack Overflow. Our experimental setting is focused on evaluating the usefulness of the extensions we proposed for PLTs, i.e., the adaptive handling of labels and online boosting.

References

1. Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *Proc. ICML 2016*, New York City, NY, USA, pages 1435–1444, 2016.
2. Nikunj C. Oza and Stuart Russell. Online bagging and boosting. In *In Artificial Intelligence and Statistics 2001*, pages 105–112. Morgan Kaufmann, 2001.
3. Gene Smith. *Tagging : people-powered metadata for the social web.* New Riders, Berkeley, CA, 2008.

In: M. Leyer (Ed.): Proceedings of the LWDA 2017 Workshops: KDML, FGWM, IR, and FGDB.
Rostock, Germany, 11.-13. September 2017, published at http://ceur-ws.org