Taxonomy-Driven Lumping for Sequence Mining[*]

Francesco Bonchi, Carlos Castillo, Debora Donato, and Aristides Gionis

Yahoo! Research
Diagonal 177, Barcelona, 080018, Spain
{bonchi,chato,debora,gionis}@yahoo-inc.com

In many application domains, events are naturally organized in a hierarchy. Whether events describe human activities, system failures, coordinates in a trajectory, or biomedical phenomena, there is often a taxonomy that should be taken into consideration. A taxonomy allow us to represent the information at a more general description level, if we choose carefully the most suitable level of granularity.

Given a taxonomy of events and a dataset of sequences of these events, we study the problem of finding efficient and effective ways to produce a compact representation of the sequences. This can be valuable by itself, or can be used to help solving other problems, such as clustering.

We model sequences with Markov models whose states correspond to nodes in the provided taxonomy, and each state represents the events in the subtree under the corresponding node. By lumping observed events to states that correspond to internal nodes in the taxonomy, we allow more compact models that are easier to understand and visualize, at the expense of a decrease in the data likelihood.

We formally define and characterize our problem, and we propose a scalable search method for finding a good trade-off between two conflicting goals: maximizing the data likelihood, and minimizing the model complexity. We implement these ideas in Taxomo, a taxonomy-driven modeler.

Taxomo receives a database of sequences of symbols, and a taxonomy over those symbols. An initial Markov model is created for the sequences without considering the taxonomy, and then refine it iteratively by merging states driven by the taxonomy. The likelihood of the data given a new model generated by this merging procedure, can be computed directly from the likelihood of the data given the model before the merging. This yields a fast model evaluation method that can explore many configurations in a short time. We also implement efficient strategies that guide the search process. We apply Taxomo in two different domains, query-log mining and mining of moving-object trajectories. The empirical evaluation confirms the feasibility and usefulness of our approach.

This is an extended abstract of an article published in the Data Mining and Knowledge Discovery journal [1].

Reference

1. Bonchi, F., Castillo, C., Donato, D., Gionia, A.: Taxonomy-driven lumping for sequence mining. Data Mining and Knowledge Discovery (2009) DOI: 10.1007/s10618-009-0141-6

[*] This is an extended abstract of an article published in the Data Mining and Knowledge Discovery Journal [1].

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, p. 29, 2009.