Label Classification in Machine Learning via Association Schemes

Katie Brodhead
Florida A&M University, Tallahassee, USA
katie.brodhead(at)famu.edu

Machine learning, as an area of computer science, attempts to learn a pattern or labeling system for given set of data, without having the rule for the data programmed ahead of time. We consider the case where some labels are provided for some portion of the data ahead time. For instance, cell phone users often label much of their data and a basic question is to learn how a user will label a new piece of data. We also admit query to a set of points for possible use in a set of training data to be used for learning and additionally allow training examples to be directly used to develop reasoning towards predicting labels of new examples.

More specifically, suppose $X = \{(x_1, L_1), (x_2, L_2), ..., (x_m, L_m)\}$ is a set of labeled data instances with data instances x_i and corresponding labels L_i. We define a non-conformity measure ν_L on X with respect to label L [3]. The idea is that a data instance has high L-non-conformity if it could easily be perceptively labeled with a non-L-label, given isn't "non-conformity" with aspects central to the property of being L. This can be put in the context of a (colored) directed graph: vertex x_i shares an (L_j-colored) edge with vertex x_j provided that $\nu_{L_j}(x_i)$ is below a certain threshold. Now, suppose that a new data point x_{m+1} from the query set is considered for possible inclusion in training data. In the parlance of statistics, suppose that a null hypothesis assumes that x_{m+1} has class label L. Following [1], we define a p-value function P_L by

$$P_L(x_{m+1}) = \frac{count\{k : \nu_L(x_k) \geq \nu_L(x_{m+1})\}}{card(X)}$$

We determine whether x_{m+1} should be selected to re-train a classifier by considering a p-value closeness matrix C whose entries $C_{ij} = |P_i - P_j|$ contain the differences of all p-values for x_{m+1} according to class labels i, j. An appeal to the Perron-Frobenius Theorem ensures that a unique maximum positive eigenvalue exists. This is used to determine a data-driven, rather than an empirically-driven, threshold for selection. Furthermore, we are able to frame results in the context of association schemes [2].

Acknowledgments. We are grateful to Akihiro Munemasa for sharing the work contained in [2].

References

[1] V. Balasubramanian, S. Chakraborty, S. Panchanathan, Generalized Query by Transduction for Online Active Learning, in: IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, 2009, 1378–1385.

[2] E. Bannai, T. Ito, *Algebraic Combinatorics I: Association Schemes*, Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA, USA (1984).

[3] V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World.* Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005).