Constrained Weighted Feature Selection

Samah Hijazi[1], Mariam Kalakech[2], Ali Kalakech[3], Denis Hamad[4]

[1,4] *Laboratoire LISIC - EA 4491 - Université du Littoral Côte d'Opale 62228 Calais Cedex, FRANCE - Email [1] : samah.hijazi@mel-etu.univ-littoral.fr - Email [4] : denis.hamad@lisic.univ-littoral.fr*

[2,3] *Faculty of Business Administration, Lebanese University, Hadath, LEBANON- Email [2] : mariam.kalakech@gmail.com - Email [3] : alikalakech@hotmail.com*

Abstract

The aim of feature selection is to identify the most informative and relevant features for a compact and accurate data representation. Generally speaking, feature selection algorithms were handled in supervised and unsupervised learning contexts. However, the semi-supervised context is more realistic where we might have only few labeled data and many others unlabeled. In this regard, another form of supervision information is available, it is based on simple pairwise comparisons [1, 2] and can be more easily obtained compared to class labels. For instance, a data pair is said to be a "must-link constraint" if its data points are similar and a "cannot-link constraint" otherwise. Recently, there was a big interest in constrained clustering that handled choosing the constraints actively and systematically, however, only few worked similarly for feature selection. Unexpectedly, [4] stated that randomly chosen constraint sets can degrade the learning performance.
Therefore, we first suggested a margin-based algorithm, Relief-Sc, for weighting features according to their data discrimination ability. It is said to find a unique relevant feature subset in a closed-form. For that, we modified ReliefF algorithm to adapt the use of cannot-link constraints with the margin concept used in [3]. We also propose to use the systematic way used by [5] to find the points that our feature selection approach might be most uncertain about, and then actively query for constraints upon these particular points. Since we can only query the oracle or expert for few constraints, we finally suggest to extend our algorithm to make use of the unlabeled data together with the chosen set of constraints. In order to validate our proposed algorithm, experiments are achieved on multiple UCI machine learning datasets and the results are prominent.

References

[1] D.Zhang, S.Chen, and Z.-H.Zhou, "Constraint Score: A new filter method for feature selection with pairwise constraints", *Pattern Recognition*, vol. 41, no. 5, pp. 1440–1451, 2008.

[2] M. Kalakech, P. Biela, L. Macaire, and D. Hamad, "Constraint scores for semi-supervised feature selection: A comparative study", *Pattern Recognition Letters*, vol. 32, no. 5, pp. 656–665, 2011.

[3] M. Yang, and J. Song, "A novel hypothesis-margin based approach for feature selection with side pairwise constraints", *Neurocomputing*, vol. 73, no. 16, pp. 2859–2872, 2010.

[4] I. Davidson, K.L. Wagstaff, and S. Basu, "Measuring constraint-set utility for partitional clustering algorithms", *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, pp. 115–126, 2006.

[5] F.L. Wauthier, N. Jojic, and M.I. Jordan, "Active spectral clustering via iterative uncertainty reduction", *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1339–1347, 2012.