From: AAAI-00 Proceedings. Copyright © 2000, AAAI (www.aaai.org). All rights reserved.

Clustering with Instance-level Constraints

Kiri Wagstaff and Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY 14850
(607) 255-5033
{wkiri,cardie}@cs.cornell.edu
http://www.cs.cornell.edu/home/wkiri/research/constraints.html

Clustering algorithms seek to discover underlying patterns in a data set automatically. To this end, they conduct a search through the space of possible organizations of the data, preferring those which group similar instances together and keep dissimilar instances apart. We claim that this search can be aided by the addition of constraints, which serve to restrict the search space and to guide the search through it.

Although clustering remains a popular area of research, to our knowledge no previous attempt has been made to incorporate hard constraints into a clustering algorithm. However, constraints have been used successfully in other unsupervised domains (e.g. interactive knowledge base construction (De Raedt, Bruynooghe, & Martens 1991)). Additionally, work has been done on incorporating general background knowledge (e.g. as a starting point in the search space (Thompson & Langley 1992) or as declarative knowledge in the form of rules on cluster membership (Talavera & Béjar 1999)).

In this work, we focus on two kind of constraints: *must-link* and *cannot-link* constraints. Both are considered hard constraints that must be satisfied (we defer an exploration of other kinds of constraints, including soft constraints, to future work). Must-link constraints specify that two instances have to be in the same cluster, while cannot-link constraints prevent two instances from being in the same cluster. We experimented with constraints using a modified version of COBWEB (Fisher 1987) that constructs a partition of the data (rather than a hierarchy). For any real-world application of this technique, constraints would be derived from problem-specific background knowledge. However, we here used randomly generated sets of constraints based on the (known but not visible to the clustering algorithm) class labels. Evaluation was done using 10-fold cross-validation with 50 random trials per fold.

We found that the incorporation of constraints can improve clustering accuracy, i.e. how close the resulting partition is to the correct partition. In experiments on four data sets (three from UCI (Blake & Merz 1998) and a part-of-speech tagging data set; $n = 50$), we saw improvements of up to 11% after incorporating 50 constraints, and up to 17% with 100 constraints. We also discovered that the type of constraint that is most effective can vary between data sets; greater increases can be obtained, for example, by using only must-link constraints with the mushroom data set. In addition, because constraints restrict the search space, we observed a corresponding decrease in runtime as more constraints were added.

We intend to apply this technique to other clustering algorithms so that they can likewise take advantage of constraint information. In addition, we plan to experiment with a variety of other data sets. In particular, some real-world domains appear likely to benefit from the inclusion of constraints, such as noun phrase coreference. In this task, background linguistic knowledge offers useful hints as to which noun phrases should be grouped together and which should not. In conclusion, we have reported on work which shows that incorporating constraints in a clustering algorithm can lead to an increase in accuracy for class discovery.

This work was supported in part by a National Science Foundation Graduate fellowship and by NSF Grant IRI-9624639. We would like to thank Westley Weimer for his suggestions as the work progressed and Peter Cheeseman, Doug Fisher, and John Stutz for email discussions.

Copyright © 2000, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

References

Blake, C. L., and Merz, C. J. 1998. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

De Raedt, L.; Bruynooghe, M.; and Martens, B. 1991. Integrity Constraints and Interactive Concept-Learning. In *Proceedings of the Eighth International Workshop on Machine Learning*, 394–398. Northwestern University, Chicago, IL: Morgan Kaufmann.

Fisher, D. 1987. Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning* 2:139–172.

Talavera, L., and Béjar, J. 1999. Integrating Declarative Knowledge in Hierarchical Clustering Tasks. In *International Symposium on Intelligent Data Analysis,* 211–222. Amsterdam, The Netherlands: Springer-Verlag.

Thompson, K., and Langley, P. 1992. Case Studies in the Use of Background Knowledge: Incremental Concept Formation. In *AAAI-92 Workshop on Constraining Learning with Prior Knowledge*, 60–68. San Mateo, CA: The AAAI Press.