Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)
Barcelona, Spain, July 13 - 14, 2015
Paper No. 328

Semi-Supervised Clustering Based on Inner Partitions Detection

Marek Śmieja, Magdalena Wiercioch
Faculty of Mathematics and Computer Science, Jagiellonian University
Lojasiewicza 6, 30-348 Krakow, Poland
marek.smieja@uj.edu.pl; magdalena.wiercioch@uj.edu.pl

Extended Abstract

Modern data engineering seeks to create clustering techniques which are flexible, scalable and can detect groups of complex shapes. It is hard to include all these requirements into one general unsupervised algorithm. Therefore, clustering methods are modifying to use the information of class labels (imposed on a sample of points) to accommodate the algorithm to particular requirements.

A lot of semi-supervised clustering algorithms incorporate hard equivalence constraints – positive constrains directly specify that two elements belong to the same group while negative constraints determine that the elements should be grouped separately. One of the most successful approach to semi-supervised clustering is a constrained EM proposed by Shental et al (2004). Equivalence constraints are used to gather points into chunklets, i.e., sets containing elements which share the same positive constraints. The algorithm fits a mixture of Gaussians to unlabeled data together with constructed chunklets. The major drawback of the constrained version of EM is that it does not handle well the situation where elements with the same positive constraint originate from several sources. This can occur when the elements with the same class label have not been generated from one simple probability distribution but for instance from the mixture of Gaussians.

We show how to partially overcome this difficulty and propose a semi-supervised clustering algorithm which deals well with the aforementioned situation. The main idea of introduced method relies on the observation that the elements with the same positive constraint can be produced by complex source, e.g. the mixture of probability distributions. Therefore, it is necessary to find their detailed description before starting the clustering of the entire data-set. To accomplish this task, the positively constrained elements are clustered individually at the initial stage of the algorithm. Individual clusterings of elements which share the same positive constraints deliver groups which are finally used for performing clustering of the entire data-set. Proposed method can also be generalized for the case when negative constraints are introduced. However, one has to pay additional attention to preserve the negative constraints during final clustering of the entire data-set.

The algorithm is capable of discovering groups with very complex structure since at the initial stage the sets of elements with the same positive constraints are split into smaller parts. This provides detailed description of the elements which share the same positive constraint. The advantages of our algorithm increase when focusing on finding partition with complex structure of each cluster.

Shental, N., Bar-Hillel, A., Hertz, T., Weinshall, D. (2004). Computing Gaussian mixture models with EM using equivalence constraints. Adv. Neural Inf. Process. Syst. 16(8), 465-472.