Semi-supervised Clustering with Constraints

Jingyuan Zhang
OSCAR Team, SSDUT, DLUT
Jun. 3, 2010

Abstract

In this presentation, I'll introduce the typical algorithms for semi-supervised clustering [1]. First, I'll introduce the generation of constraints for datasets. The basic instance level constraints mainly are must-link and cannot-link. Then I'll talk about using constraints to aid and bias clustering in detail. In this section, several clustering algorithms will be explained, including the typical unsupervised K-means algorithm and the COP-kmeans [2], the distance metric learning techniques and a probabilistic framework using HMRF (Hidden Markov Random Field) model. In these algorithms, COP-kmeans incorporates conjunctions of constraints to k-means algorithm and replaces must-linked instances by the average of connected components. However, in some cases, the constraints would contradict each other and the constraints would be over-used by COP-kmeans, which can result in bad clustering performance. Thus, using constraints to learn distance function was proposed in [3, 4]. In these techniques, must linked data tend to have shorter distances from each other while cannot linked data tend to be far away from each other. However, it is not suitable for spatial datasets and a probabilistic framework was designed in [5] to handle this problem. I'll introduce you the basic concept of Markov Chain (MC), Markov Random Field (MRF), Markov Model (MM) and Hidden Markov Model (HMM) with specific examples and analogies in daily life. Besides, the Expectation-Maximization algorithm (EM) will also be explained in a similar way.

References

[1] I. Davidson, S. Basu. A Survey of Clustering with Instance Level Constraints. In KDD 2007.
[2] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, Constrained K-means Clustering with Background Knowledge. In ICML 2001.
[3] D. Klein, S. D. Kamvar and C. D. Manning. From Instance-Level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering. In ICML 2002.
[4] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. NIPS 15, 2003
[5] S. Basu, M. Bilenko and R. J. Mooney. A Probabilistic Framework for Semi-Supervised Clustering. In KDD 2004.