Learning Kernels for Semi-Supervised Clustering

Bojun Yan
George Mason University, USA

Carlotta Domeniconi
George Mason University, USA

INTRODUCTION

As a recent emerging technique, semi-supervised clustering has attracted significant research interest. Compared to traditional clustering algorithms, which only use unlabeled data, semi-supervised clustering employs both unlabeled and supervised data to obtain a partitioning that conforms more closely to the user's preferences. Several recent papers have discussed this problem (Cohn, Caruana, & McCallum, 2003; Bar-Hillel, Hertz, Shental, & Weinshall, 2003; Xing, Ng, Jordan, & Russell, 2003; Basu, Bilenko, & Mooney, 2004; Kulis, Dhillon, & Mooney, 2005).

In semi-supervised clustering, limited supervision is provided as input. The supervision can have the form of labeled data or pairwise constraints. In many applications it is natural to assume that pairwise constraints are available (Bar-Hillel, Hertz, Shental, & Weinshall, 2003; Wagstaff, Cardie, Rogers, & Schroedl, 2001). For example, in protein interaction and gene expression data (Segal, Wang, & Koller, 2003), pairwise constraints can be derived from the background domain knowledge. Similarly, in information and image retrieval, it is easy for the user to provide feedback concerning a qualitative measure of similarity or dissimilarity between pairs of objects. Thus, in these cases, although class labels may be unknown, a user can still specify whether pairs of points belong to the same cluster (Must-Link) or to different ones (Cannot-Link). Furthermore, a set of classified points implies an equivalent set of pairwise constraints, but not vice versa. Recently, a kernel method for semi-supervised clustering has been introduced (Kulis, Dhillon, & Mooney, 2005). This technique extends semi-supervised clustering to a kernel space, thus enabling the discovery of clusters with non-linear boundaries in input space. While a powerful technique, the applicability of a kernel-based semi-supervised clustering approach is limited in practice, due to the critical settings of kernel's parameters. In fact, the chosen parameter values can largely affect the quality of the results. While solutions have been proposed in supervised learning to estimate the optimal kernel's parameters, the problem presents open challenges when no labeled data are provided, and all we have available is a set of pairwise constraints.

BACKGROUND

In the context of supervised learning, the work in (Chapelle & Vapnik) considers the problem of automatically tuning multiple parameters for a support vector machine. This is achieved by minimizing the estimated generalization error achieved by means of a gradient descent approach over the set of parameters. In (Wang, Xu, Lu, & Zhang, 2002), a Fisher discriminant rule is used to estimate the optimal spread parameter of a Gaussian kernel. The authors in (Huang, Yuen, Chen & Lai, 2004) propose a new criterion to address the selection of kernel's parameters within a kernel Fisher discriminant analysis framework for face recognition. A new formulation is derived to optimize the parameters of a Gaussian kernel based on a gradient descent algorithm. This research makes use of labeled data to address classification problems. In contrast, the approach we discuss in this chapter optimizes kernel's parameters based on unlabeled data and pairwise constraints, and aims at solving clustering problems. In the context of semi-supervised clustering, (Cohn, Caruana, & McCallum, 2003) uses gradient descent combined with a weighted Jensen-Shannon divergence for EM clustering. (Bar-Hillel, Hertz, Shental, & Weinshall, 2003) proposes a Redundant Component Analysis (RCA) algorithm that uses only must-link constraints to learn a Mahalanobis distance. (Xing, Ng, Jordan, & Russell, 2003) utilizes both must-link and cannot-link constraints.
Related Content

Data Mining Methods for Microarray Data Analysis
Lei Yu and Huan Liu (2005). Encyclopedia of Data Warehousing and Mining (pp. 283-287).
www.igi-global.com/chapter/data-mining-methods-microarray-data/10608?camid=4v1a

Bioinformatics Data Management and Data Mining
www.igi-global.com/chapter/bioinformatics-data-management-data-mining/7727?camid=4v1a

Genetic Programming for Automatically Constructing Data Mining Algorithms
www.igi-global.com/chapter/genetic-programming-automatically-constructing-data/10932?camid=4v1a

Automatic Musical Instrument Sound Classification
www.igi-global.com/chapter/automatic-muscial-instrument-sound-classification/10571?camid=4v1a