Constrained Spectral Clustering Using Absorbing Markov Chains

Jianyuan Li[1,2] and Jihong Guan[1]

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[2] School of Mathematics and Computer Science, Shanxi Normal University, Linfen 041000, China
jylbob@gmail.com, jhguan@tongji.edu.cn

Abstract. Constrained spectral clustering (CSC) has recently shown great promise in improving clustering accuracy or catering for some specific grouping bias by encoding pairwise constraints into spectral clustering. Essentially, the existing CSC algorithms coarsely lie in two camps in terms of encoding pairwise constraints: (1) they modify the original similarity matrix to encode pairwise constraints; (2) they regularize the spectral embedding to encode pairwise constraints. Those methods have made significant progresses, but little of them takes the extensional sense of pairwise constraints into account, e.g., respective neighbors of two musk-link points lie in a same cluster with certain high probabilities, and respective neighbors of two cannot-link points lie in different clusters with certain high probabilities, etc. In this paper, we use absorbing Markov chains to formulate the extensional sense of instance-level constraints as such, under the assumption that the formulation aids in improving the accuracy of CSC. We describe a new CSC algorithm which could propagates the extensional sense over a partly-labeled affinity graph. Experiments over publicly available datasets verify the performance of our algorithm.

1 Introduction

Clustering is an important unsupervised learning method that aims to detect structures in vector data, finding groups in which patterns of vector data are as similar as possible [1], or finding groups that statistically form smooth submanifolds. However, purely unsupervised learning often can not satisfy some specific grouping bias, since certain features might play noisy (or redundant) roles to the grouping bias. Motivated by this, researchers have explored a class of methods that could incorporate accessible supervision with clustering algorithms for improving clustering accuracy, as reported in [2–10].

The most accessible supervision is called pairwise constraints, meaning whether pairwise instances lie in a same group or not, and so the problem in question is called *constrained clustering*. To incorporate pairwise constraints with clustering is often practical and useful in many fields, e.g., in document clustering, whether two documents concern about a same topic or not, can be readily judged through skimming;

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 52–63, 2012.

Table 1. Notation description

Symbols	Description
X	a vector dataset $\{\mathbf{x}_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$
d	dimensionality of X
n	cardinality of X
m	the number of constrained points
G	the affinity graph over X
$\mathbf{W}$	the affinity matrix over X
$\mathbf{P}$	the original probability transition matrix
$\mathbf{F}$	the fundamental matrix of absorbing Markov chains
$\mathbf{I}$	the diagonal matrix with diagonal elements ones
$\mathcal{M}$	the set of Must-Link constraint pairs
$\mathcal{C}$	the set of Cannot-Link constraint pairs
$\mathbf{Y}$	the m-by-m constraint matrix

and in image segmentation, whether two patches represent a same meaningful entity or not, can be easily judged through observing.

The key problem to be solved is how to encode pairwise constraints into an existing clustering algorithm. In this paper, we focus on how to encode pairwise constraints into *spectral clustering*, since spectral clustering has proved to be competitive in terms of the accuracy and the wide applicability [11–17]. Existing work about this problem can be coarsely broken into two camps: (1) they modify the original similarity matrix to encode pairwise constraints, such as [5, 6, 10, 18]; (2) they regularize the spectral embedding to encode pairwise constraints, such as [8]. However, little work has ever explicitly considered the extensional sense of pairwise constraints. The aim of this paper is to explicitly characterize the extensional sense and to improve clustering accuracy. In essence, our work belongs to the first camp, using the extensional sense to modify the original affinity matrix.

The structure of this paper is introduced below. In Section 2, we review the existing constrained clustering algorithms. In Section 3 we interpret the pairwise constraints and its extensional sense, and we further characterize the extensional sense based on the theory of absorbing Markov chains. In Section 4, we describe our constrained spectral clustering algorithm. In Section 5, we evaluate the proposed algorithm over benchmark datasets. In Section 6, we draw the conclusion and mention the future work.

The notation that we used is in general as follows. Vectors are denoted by bold lower-case letters. Matrices are denoted by upper-case ones or calligraphy letters. Sets are denoted by italic upper-case letters or calligraphy letters. Scalars are denoted by italic lower-case letters. A part of symbols are introduced in Table 1.

2 Related Work

Constrained spectral clustering focuses on mitigating the blindness of unsupervised spectral clustering by incorporating few user supervision. By now, there