Semi-supervised K-Way Spectral Clustering with Determination of Number of Clusters

Guillaume Wacquet, Émilie Poisson-Caillault, and Pierre-Alexandre Hébert

LISIC - Lab. of Computing, Signal and Image Processing in Côte d'Opale,
Université Lille Nord de France, ULCO, Calais, France
{Guillaume.Wacquet,Emilie.Caillault,
Pierre-Alexandre.Hebert}@lisic.univ-littoral.fr
http://www-lisic.univ-littoral.fr

Abstract. In this paper, we propose a new K-way semi-supervised spectral clustering method able to estimate the number of clusters automatically and then to integrate some limited supervisory information. Indeed, spectral clustering can be guided thanks to the provision of prior knowledge. For the automatic determination of the number of clusters, we propose to use a criterion based on an outlier number minimization. Then, the prior knowledge consists of pairwise constraints which indicate whether a pair of objects belongs to a same cluster (*Must-Link* constraints) or not (*Cannot-Link* constraints). The spectral clustering then aims at optimizing a cost function built as a classical *Multiple Normalized Cut* measure, modified in order to penalize the non-respect of these constraints. We show the relevance of the proposed method with some UCI datasets. For experiments, a comparison with other semi-supervised clustering algorithms using pairwise constraints is proposed.

Keywords: Spectral embedding, Within-cluster cohesion, Semi-supervised clustering, Pairwise constraints.

1 Introduction

The proposed semi-supervised clustering methodology aims at clustering unknown data, considering a real case, when some expert can add some knowledge. More precisely, it is composed of two main steps: first, the build of an initial convenient clustering, without any knowledge, which is then theoritically used by an expert to assess or correct some clustering results, using pairwise constraints. In a second step, a semi-supervised clustering is then used to adjust the initial clustering, by integrating the newly available knowledge.

Among the whole set of clustering methods, we focus on algorithms able to generate discriminant representations, which can then be clustered by simple algorithms, like K-means. We look for a subspace conjointly maximizing within-cluster cohesion and between-clusters separation. Both measures can be gathered in one criterion like the Multiple Normalized Graph Cut ($MNCut$) [7]. This criterion is the basis of spectral clustering algorithms, "in vogue" in the literature, thanks to their effective global optimization and their simplicity of implementation. Both these advantages are due to the

K. Madani et al. (Eds.): *Computational Intelligence*, SCI 465, pp. 317–332.
DOI: 10.1007/978-3-642-35638-4_21

main step: the eigenvectors extraction from a similarity matrix computed on the dataset [13][4]. Similarity matrix gathers the complete information used by the method, telling for each pair of objects how close they are. Moreover, spectral clustering algorithms are able to deal with complex cases including "non-globular" or non-linearly separable clusters.

As the first step, we propose a methodology to build an optimal partition, using Ng's algorithm [5] which tends to produce particularly well discriminative representation space, to estimate the number of clusters without any kind of knowledges. This determination of number K is based on a cluster representativeness, defined as the proportion of outliers. Indeed, the only $MNCut$ value cannot always guarantee a good partition because of its minimization process. This is the reason why we introduce two additional criteria: the limitation of outliers and the minimization of the number of clusters.

As the second step, the method comes within a real context, where the obtained partition is presented to experts which can validate it and can give some additional informations. Indeed, in recent years, methods incorporating prior knowledge in their clustering process have emerged as both relevant and effective in several applications, such as image segmentation [4], information retrieval or document analysis [2]. The prior knowledge is generally provided in two forms: class labels, and pairwise constraints. Labelling data is a hard and long task. Pairwise constraints simply indicate if two instances must be in the same cluster (*Must-Link*) or not (*Cannot-Link*). They are easier to collect from experts than labels [11]. In this work, pairwise constraints are randomly built from ground-truth labels. Then, we assume that the generated knowledges are true and relevant.

In this paper, we propose a new algorithm able to integrate constraints in the multiclass spectral clustering process, using a penalty term. The proposed method aims at minimizing the Multiple Normalized Cut criterion, while penalizing the non-respect of the given set of constraints. Moreover, a convenient weight, easily interpretable, is introduced in order to balance the $MNCut$ and the penalty term, i.e. the impact of the original data structure and the contribution of the constraints.

The paper is organized into four sections. The first one is theoretical and introduces some basic notations and spectral clustering methods of the literature. In a second section, we present the first step of the proposed method, i.e. a method based on a measure of the representativeness cluster allowing to estimate the number of clusters automatically. The third section presents the second step, i.e. the proposed semi-supervised K-way spectral clustering method able to integrate some prior knowledges. The last section assesses the performances of our method versus some semi-supervised algorithms of the literature on public databases extracted from UCI repository[1]. The results are finally presented, for different proportions of known constrained pairs.

2 Graph Embedding and Spectral Clustering

Spectral clustering is generally considered as a clustering method aiming at minimizing a *Normalized Cut* criterion between $K = 2$ clusters ($NCut$), or a *Multiple Normalized Cut* between $K \geq 2$ clusters ($MNCut$) [4][5][7]. The first measure, $NCut$, assesses

[1] `http://archive.ics.uci.edu/ml/`