Pattern Recognition 47 (2014) 820–832

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Semi-supervised clustering via multi-level random walk

Ping He, Xiaohua Xu*, Kongfa Hu, Ling Chen

Department of Computer Science, Yangzhou University, Yangzhou 225009, China

ARTICLE INFO

Article history:
Received 12 February 2013
Received in revised form
17 June 2013
Accepted 31 July 2013
Available online 12 August 2013

Keywords:
Semi-supervised clustering
Pairwise constraint
Influence expansion
Multi-level random walk
Spectral clustering

ABSTRACT

A key issue of semi-supervised clustering is how to utilize the limited but informative pairwise constraints. In this paper, we propose a new graph-based constrained clustering algorithm, named SCRAWL. It is composed of two random walks with different granularities. In the lower-level random walk, SCRAWL partitions the vertices (i.e., data points) into constrained and unconstrained ones, according to whether they are in the pairwise constraints. For every constrained vertex, its influence range, or the degrees of influence it exerts on the unconstrained vertices, is encapsulated in an intermediate structure called component. The edge set between each pair of components determines the affecting scope of the pairwise constraints. In the higher-level random walk, SCRAWL enforces the pairwise constraints on the components, so that the constraint influence can be propagated to the unconstrained edges. At last, we combine the cluster membership of all the components to obtain the cluster assignment for each vertex. The promising experimental results on both synthetic and real-world data sets demonstrate the effectiveness of our method.

1. Introduction

Semi-supervised clustering, also called constrained clustering, has become a hotspot in the current research of machine learning and data mining communities. Compared with traditional clustering [1], semi-supervised clustering takes advantage of the additional prior knowledge, such as cluster seeds or pairwise constraints, to improve the clustering result and avoid clustering ambiguity.

There are two types of supervision mostly used in semi-supervised clustering. The first one is cluster seed set [2], very similar to the labeled data set in semi-supervised classification. The second type is pairwise constraint set, which specifies the pairs of data belonging to the same cluster (must-link constraints) or different clusters (cannot-link constraints) [3]. If we view every data point as a vertex on graph, then the first category of semi-supervised clustering is a *vertex-constrained* learning problem, while the second category is as an *edge-constrained* learning problem. Since the edge constraints can be inferred from the vertex constraints, but not vice versa, it is more challenging to deal with the edge-constrained clustering problems than those vertex-constrained ones.

By far, various methods have been proposed to handle semi-supervised clustering with pairwise constraints. Generally, they can be classified into two lines. The first line, namely metric learning, learns optimized metric(s) to keep the must-linked data close and the cannot-linked data far away [4–6]. Most of the existing metric learning algorithms learn linear Mahalanobis distance metrics [4,5], but Wu et al. [6] develop a novel scheme to learn nonlinear Bregman distance functions. However, they have a generally known disadvantage that metric learning approaches require a large number of pairwise constraints to learn the correct metrics [7]. Moreover, they heavily rely on the prior assumption about the metric scope, which is hard to predict in advance. For instance, Xing et al. [4] assume that all the data points share a single global metric, while Bilenko et al. [8] assume that every cluster has an independent local metric.

The second line of semi-supervised clustering algorithms focuses on adapting the existing clustering (generative) or classification (discriminative) models to deal with this problem. The early algorithms adapt the traditional methods like k-means [9], all-pairs shortest path [10], and Gaussian mixtures models [11] to find a clustering result that can satisfy all the pairwise constraints greedily. However, without the mechanism of backtracking, they may fail to find a satisfying partition even when there exists one. To tackle the sub-optimality problem, people adopt bio-inspired metaheuristic methods, such as genetic algorithm [12] and Ant Colony Optimization [13], which can explore the solution space more exhaustively and hence have a larger chance to find the global optimal solution. In recent years, more and more graph-based methods are incorporated in semi-supervised clustering algorithms. Lu [14] generalize the MAP Gaussian process classifiers to express the uncertainty information associated with the pairwise constraints in a probabilistic framework. In addition, semi-supervised clustering based on kernel methods [15], maximum

* Corresponding author. Tel.: +86 514 879 78309; fax: +86 514 878 87937.
E-mail address: arterx@gmail.com (X. Xu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.023

margin clustering [7], ensembles [16] and fuzzy *c*-means [17] have all been developed along this line.

Along the second line, there is an emerging trend in developing semi-supervised clustering algorithms based on the spectral method [18]. Kamvar et al. [19] first modify the pairwise similarity matrix by setting the must-link similarities as 1 and the cannot-link similarities as 0, then apply spectral clustering on the modified similarity matrix. However, the 1/0 modification strategy seems extreme because the data in the same cluster may not coincide and the data in different clusters probably share similar attributes. To overcome the shortcoming, Kulis et al. [20] propose a reward/penalty strategy, which adds a reward to the must-link similarities and subtracts a penalty from the cannot-link similarities. The drawback, as Li et al. [21] criticized, is that it may cause non-positive-semidefinite problem for convergence if the penalty is larger than the original similarities.

It is soon realized that by only revising the similarities of the constrained edges, it is hard to utilize the limited but informative pairwise constraints. A straightforward solution is to expand the constraint influence to the unconstrained edges, but the key issue lies in how. Although diverse efforts have been made, including the formulation of the constrained normalized cut [22], the alteration of the Laplacian matrix eigenspace [23], and the incorporation of the Gaussian process [24], they either cannot deal with the multi-class semi-supervised clustering problems or fail to handle the cannot-link constraints. Wang and Davidson [25] develop an objective function that allows real-valued degree-of-belief constraints, but it can hardly produce satisfactory result when the number of pairwise constraints is small. Li et al. [21] combine the spectral method with global metric learning to adapt the spectral embedding of the data as consistent with the pairwise constraints as possible. Nevertheless, a metric is rarely uniform in the whole domain. In another word, the structure of patterns may vary between different local neighborhoods. Thus a more appropriate way is to spread the pairwise constraints locally and exert greater influence on the nearby edges than on the faraway edges.

To confine the influence of the pairwise constraints to local areas, it is a natural choice to replace a global metric with several local metrics. Bilenko et al. [8] integrate local metric learning with constrained *k* means to learn an individual local metric for every cluster. The disadvantage is that they cannot deal with the data sets containing two or more local metrics in one cluster. Moreover, users need to provide much more pairwise constraints to ensure the correctness of all the local metrics. Besides, Lu and Peng [26] transform the pairwise constraint propagation into solving a continuous-time Lyapunov equation, which requires a high computational cost. Although the authors provide an approximation strategy to obtain a suboptimal solution, it still costs quadratic time complexity.

In this paper, we propose a novel approach to spreading the constraint influence to the surrounding unconstrained edges with sufficient smoothness. To this end, we decompose the constraint propagation process into three steps. First, we extract the vertices in the pairwise constraints, called constrained vertices (edge → vertex). Second, we determine the influence range of every constrained vertex by computing the degrees of influence it exerts on the unconstrained vertices (vertex → vertex). Third, we derive the affecting scope of each pairwise constraint, and enforce the pairwise constraints on the affected edges. During these three steps, each pairwise constraint is at first treated as a single constrained edge, then transformed into the influence range of two constrained vertices, and at last expanded to a group of affected edges. Therefore, we call this procedure an “edge → vertex → edge” constraint utilization strategy.

More specifically, our algorithm named SCRAWL, short for Semi-supervised Clustering via RAndom WaLk, is composed of two random walks with different granularities. In the lower-level random walk, SCRAWL partitions the vertex set into the constrained and unconstrained two vertex subsets. Then it determines the influence range of every constrained vertex by translating the problem to a well-studied issue in semi-supervised classification, that is estimating the probabilities of the unlabeled data belonging to the same class of a labeled data [27]. For this purpose, a semi-supervised classification algorithm, label propagation [28], is incorporated in SCRAWL. We further encapsulate the vertices within the influence range of every constrained vertex in an intermediate structure called component. The component membership degree of each vertex equals to the degree of influence it receives from the constrained vertex. Since the overall component membership degree of each vertex is 1, we divide a whole vertex to multiple fractional vertices [29], e.g., $v=\left[\frac{1}{2}v,\frac{1}{3}v,\frac{1}{6}v\right]$, according to its degrees of different component membership. In this point of view, a component is the union of the fractional vertices affected by a distinct constrained vertex. In the higher-level random walk, SCRAWL derives the affecting scope of each pairwise constraint, which is the edge set connecting the components around the two constrained vertices. We call such an edge between two fractional vertices in different components, e.g. $\langle\frac{1}{2}v_i,\frac{1}{3}v_j\rangle_{(1/2)v_i\,\in\,\text{component}_1,(1/3)v_j\,\in\,\text{component}_2}=\frac{1}{6}\langle v_i,v_j\rangle$, a fractional edge. Its fraction, determined by the product of the fractions of the connected vertices (e.g., $\frac{1}{6}=\frac{1}{2}\cdot\frac{1}{3}$), indicates the degree of influence that the whole edge ($\langle v_i,v_j\rangle$) receives from the constrained edge. To expand the constraint influence, we enforce the pairwise constraints on the fractional edges among components, and group the components into different clusters. Finally, we obtain the cluster assignment for each vertex by combining the cluster membership of the fractional vertices distributed in different components. The promising experimental results on the synthetic data sets, UCI database and image segmentations demonstrate the effectiveness of SCRAWL.

There are several aspects of our proposed approach worthwhile to highlight here:

- SCRAWL can propagate the pairwise constraints to the surrounding unconstrained edges in proportion to the degrees of influence they receive from the constrained edges. The greater influence an unconstrained edge receives from a constrained edge, the more likely it is to satisfy the same pairwise constraint.
- The existing graph-based semi-supervised clustering algorithms confine the utilization of the pairwise constraints on edges. In contrast, SCRAWL develops an “edge → vertex → edge” constraint utilization strategy, which can expand a single constrained edge, through its two connected vertices, to a group of affected edges.
- SCRAWL introduces an intermediate structure between the fine-grained vertex and the coarse-grained cluster, called “component”. It can effectively uncover the underlying substructures of the clusters.
- SCRAWL establishes a connection between semi-supervised clustering and semi-supervised classification algorithms. It provides a new way to develop semi-supervised clustering algorithms based on the semi-supervised classification algorithms, which can predict the degrees of different class membership for each unlabeled data.
- SCRAWL can effectively handle the clustering problems with extremely small or large amount of pairwise constraints.
- For large real-world data sets, the time complexity of SCRAWL is approximately linear, if given a *k* NN sparse similarity matrix.

The remainder of this paper is organized as follows. Section 2 introduces the label propagation algorithm incorporated in SCRAWL. Section 3 describes the algorithm of SCRAWL in detail. Section 4 discusses the parameters of SCRAWL. Section 5 evaluates