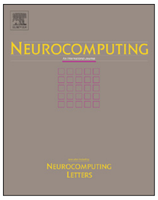
Neurocomputing 188 (2016) 239–252

Contents lists available at ScienceDirect

Neurocomputing

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



Improving clustering with constrained communities

Xiaohua Xu [a,b,c], Ping He [a,b,c,*]

[a] Department of Computer Science, Yangzhou University, Yangzhou, 225009, China
[b] Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, 225009, China
[c] Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou, 225009, China

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form
17 September 2014
Accepted 29 September 2014
Available online 2 December 2015

Keywords:
Constrained clustering
Constrained community
Transductive learning
Community discovery

ABSTRACT

In this paper, we propose a new constrained clustering algorithm, named Constrained Community Clustering (C^3). It can utilize both must-link and cannot-link constraints that specify the pairs of data that belong to the same or different clusters. Instead of directly enforcing the pairwise constraints on the constrained data, C^3 first builds constrained communities around each constrained data, and then exert pairwise constraints on the constrained communities. Therefore, C^3 can not only extend the influence of pairwise constraints to the surrounding unconstrained data, but also uncover the underlying substructures of the clusters. The promising experimental results on the real-world text documents, handwritten digits, alphabetic characters, face recognition, and community discovery illustrate the effectiveness of our method.

1. Introduction

Clustering used to be defined as an unsupervised learning problem, whose goal is to organize a collection of data into different groups, so that the within-group data is more densely connected than those distributed in different groups. This situation lasts until cluster seeds [1] and pairwise constraints [2] are introduced to guide the clustering process. We call the clustering with pairwise constraints as constrained clustering. Compared with traditional unsupervised clustering, where multiple reasonable partitions can be applied to one problem, constrained clustering is equipped with more prior knowledge and hence directed to a more specific partition. By far, constrained clustering has been applied to a variety of applications such as discovery of disease trajectories [3], analysis of gene expression data [4], and privacy-preserving data publishing [5].

Constrained clustering typically deals with two types of pairwise constraints, namely must-links and cannot-links. Must-link constraints specify the pairs of data that belong to the same cluster, while cannot-link constraints specify the pairs of data that belong to different clusters. With must-link and/or cannot-link constraints, constrained clustering aims to find out the user-desired clustering result indicated in an implicit way.

There have been various constrained clustering algorithms proposed so far. A major group of constrained clustering algorithms are derived from unsupervised methods, including *k*-means [2], all-pairs shortest path [6], Gaussian mixtures models [7], Gaussian process [3,8] and Spectral clustering [9–12]. Among them, Wagstaff et al. [2] proposed the first constrained clustering algorithm that adapts the traditional *k*-means algorithm to find a solution that satisfy all the pairwise constraints. One year later, Klein et al. [6] interpreted qualitative pairwise constraints as quantitative distances (or proximities) to infer all-pairs shortest path. Generative models, which are widely used in unsupervised clustering, are also adopted in constrained clustering algorithms. Shental et al. [7] incorporated pairwise constraints into the density estimation of Gaussian mixture models, while Ross et al. [3] used pairwise constraints to learn a variational Dirichlet process mixture of Gaussian processes. In recent days, Calandriello et al. [13] extended the unsupervised information-maximization clustering algorithm to deal with constrained clustering, but the proposed algorithm cannot appropriately deal with cannot-link constraints in multi-class problem. With the popularity of spectral method [14], constrained spectral clustering now becomes an emerging trend in the development of constrained clustering algorithms. Kamvar et al. [9] proposed the first constrained spectral clustering algorithm that applies spectral clustering to the modified similarity matrix, where the must-link similarities are set 1 and cannot-link similarities are set 0. Kulis et al. [11] adopted the similar paradigm as [9], but only employed another similarity modification strategy, i.e. added a reward to must-link similarities and subtracted a penalty to cannot-link similarities. The disadvantage of these two works lies in that they confine the influence of pairwise constraints to the constrained data. To overcome the shortcoming, efforts have been made to integrate the

[*] Corresponding author. Tel.: +8651487978309; fax: +8651487887937.
E-mail addresses: angeletx@gmail.com, arterx@gmail.com (P. He).

http://dx.doi.org/10.1016/j.neucom.2014.09.106
0925-2312/© 2015 Elsevier B.V. All rights reserved.

satisfaction of pairwise constraints into the optimization target of constrained spectral clustering, so that sparse pairwise constraints can be propagated to the unlabeled data. Yu et al. [15] defined a constrained normalized cut criterion that combines clustering smoothness with fairness on labeled data. Rangapuram et al. [12] provided a tight relaxation of constrained normalized cut and guarantee to satisfy all the pairwise constraints. However, Yu et al. [15] failed to deal with cannot-link constraints, while Rangapuram et al. [12] failed to handle multi-class problems.

Another group of constrained clustering algorithms incorporates metric learning to transform the clustering data into a new feature space, so that the must-linked data are mapped nearby while the cannot-linked data are mapped faraway [16–19]. Global metric learning is usually preferred for the simplicity of its model. Xing et al. [16] learned a global Mahalanobis metric to spread the constraint influence to all the data points uniformly. Instead of performing metric learning in the original feature space [16], Li et al. [20] learned a global metric in the low-dimensional spectral space, while Anand et al. [21] learned a global metric in the high-dimensional kernel space for constrained clustering. In order to handle high-dimensional data efficiently, Wu et al. [22] proposed to learn a nonlinear Bregman distance function with pairwise constraints. Besides, Alzate et al. [23] learned a modified kernel matrix by incorporating hard or soft pairwise constraints into spectral clustering with natural out-of-sample extension. However, in the real-world applications, a dataset rarely has only one uniform metric. To make up for this deficiency, Bilenko et al. [24] learned a local metric for each cluster to keep the constraint influence within the related clusters. Similarly, Wang et al. [25] chose a fixed radius to determine the neighborhood of pairwise constraints before local metric learning. However, none of the prior assumptions about the scope of metric, no matter global, local or a circle with fixed radius, can universally hold true for all the datasets. A more appropriate way is to adaptively determine the influence range of each pairwise constraint.

In this paper, we propose a Constrained Community Clustering (C^3) algorithm, which can propagate the influence of pairwise constraints to the unconstrained data in proportion to the affinities among constrained and unconstrained data. To this end, we extend a pairwise constraint to the relationship between two constrained communities, where each constrained community is defined as a group of data under the influence of a constrained data. The advantage of introducing constrained community lies in that it not only adaptively determines the influence range of each pairwise constraint, but also reveals the underlying sub-structures of the clusters. Besides, similar to the communities in social network, constrained communities allow overlapping. We enforce every pairwise constraint on the constrained communities, so that the constraint influence can be propagated to the unconstrained data. At last, the expanded influence of different pairwise constraints are integrated into a cluster indicating matrix for cluster assignment. Sufficient experiments with comprehensive evaluation criteria demonstrate the superiority of C^3 over other related constrained clustering algorithms on text documents, handwritten digits, alphabetic characters, face recognition, and community discovery applications. Besides, the impact of the parameters on the clustering performance of C^3 is also analysed.

It is worthwhile to highlight several aspects of the proposed approach as follows.

- C^3 determines the range of each constrained community by taking a balance between the spreading community indicating matrix and the initial community assignment.
- We propose an exponential adaptation method to adapt the pairwise affinity matrix and the community affinity matrix, which are both symmetrically normalized, to be consistent with the pairwise constraints. It ensures that, after the modification, the elements in the adapted affinity matrices are still within the range of [0,1].
- The time complexity of C^3 is smaller than traditional graph-based methods. The time complexity of C^3 is $O(n)$ when given the precomputed affinity matrix, while the traditional methods costs $O(n^3)$ for the eigenvalue decomposition of the full affinity matrix.
- The space complexity of C^3 is smaller than traditional graph-based methods. The storage requirement of C^3 is $O(kn)$, where k is the size of neighborhood (or the sparsity of affinity matrix) and n is the number of data instances, while the traditional methods require $O(n^2)$ to store the full affinity matrix.

The remainder of this paper is organized as follows. Section 2 provides preliminaries about transductive learning and its relationship with C^3 algorithm. Section 3 describes the detailed procedures of C^3. Section 4 evaluates the clustering performance of C^3 in comparison with other state-of-the-art methods, and Section 5 concludes the whole paper.

2. Preliminaries

Transductive learning [26] refers to the following problem: given a dataset $\mathcal{D}=(\mathcal{X}_l,\mathcal{Y}_l,\mathcal{X}_u)$, where $\mathcal{X}_l=\{x_1,\ldots,x_l\}$ is the labeled data subset, $\mathcal{Y}_l=\{y_1,\ldots,y_l\}$ is the label set of $\mathcal{X}_l$, $y_i\in\{c_1,\ldots,c_p\}$, p is the number of classes, $\mathcal{X}_u=\{x_{l+1},\ldots,x_n\}$ is the unlabeled data subset, the goal is to predict the label set of $\mathcal{X}_u$, denoted as $\mathcal{Y}_u$. Transductive learning is related with our C^3 algorithm in building constrained communities.

Let $A=(a_{ij})_{n\times n}$ denote the affinity matrix of $\mathcal{X}=\mathcal{X}_l\cup\mathcal{X}_u$,

$$a_{ij}=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{1}$$

S denote the normalized affinity matrix,

$$S=D^{-1/2}AD^{-1/2} \tag{2}$$

where $D=(d_{ii})_{n\times n}$ is the diagonal degree matrix with $d_{ii}=\sum_j a_{ij}$. The estimated labels of both labeled and unlabeled data are recorded in a label indicating matrix $\hat{Y}=[\hat{Y}_l^T\hat{Y}_u^T]^T$, where $\hat{Y}_l$ is the $l\times c$ submatrix indicating the known labels of $\mathcal{Y}_l$, $\hat{Y}_u$ is the $(n-l)\times c$ submatrix indicating the predicted labels of $\mathcal{Y}_u$.

Among various transductive learning algorithms, Zhou et al. [27] spread the known labels via keeping both global and local consistency. In their algorithm, the computation of $\hat{Y}$ is performed in an iterative manner:

$$\hat{Y}^{t+1}=\alpha S\hat{Y}^t+(1-\alpha)\hat{Y}^0 \tag{3}$$

where $\hat{Y}^0$ is the initial value of $\hat{Y}$, $\alpha\in(0,1]$ is a tradeoff parameter between the spreading of $\hat{Y}^t$ and $\hat{Y}^0$. It has been proved [27] that Eq. (3) converges to a simple solution:

$$\hat{Y}^{\infty}=(1-\alpha)(I-\alpha S)^{-1}\hat{Y}^0 \tag{4}$$

The submatrix of $\hat{Y}_u^{\infty}$ is used to assign the unlabeled data with the most probable labels, $y_i=\arg\max_j\hat{y}_{ij}$, $\forall\ x_i\in\mathcal{X}_u$.

3. Constrained Community Clustering

Constrained clustering can be formulated as follows. Given an unlabeled dataset $\mathcal{X}$ and a pairwise constraint set $\mathcal{C}=C_=\cup C_{\neq}$, $\mathcal{C}_=$ is the must-link constraint subset, $\mathcal{C}_{\neq}$ is the cannot-link constraint subset. The must-link constraint $(x_i,x_j,=)\in\mathcal{C}_=$