Applied Soft Computing 11 (2011) 5281–5291

Contents lists available at ScienceDirect

Applied Soft Computing

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

A new semi-supervised clustering algorithm with pairwise constraints by competitive agglomeration

Cui-Fang Gao[a,b], Xiao-Jun Wu[b,*]

[a] School of Science, Jiangnan University, Wuxi 214122, PR China
[b] School of Computer Science and Technology, Jiangnan University, Wuxi 214122, PR China

ARTICLE INFO

Article history:
Received 15 April 2009
Received in revised form 27 October 2010
Accepted 15 May 2011
Available online 20 May 2011

Keywords:
Fuzzy clustering
Semi-supervised
Pairwise constraints
Penalty cost function

ABSTRACT

Recently semi-supervised fuzzy clustering with pairwise constraints was developed, in which the disagreement on the magnitude order between penalty cost function and the basic objective function will cause over adjustment of membership values and their deviation from the normal range. In order to solve this problem, an improved semi-supervised fuzzy clustering algorithm with pairwise constraints (SCAPC) was proposed based on a redefined objective function. The new penalty cost function in SCAPC theoretically conforms to the methodology of classical fuzzy clustering, which is expressed as the violation cost incurred by the pairs, and has the same magnitude order as the basic objective function. Experimental results on benchmark datasets and images showed that SCAPC can produce more accurate clustering by moderately enhancing or reducing the ambiguous memberships. Research indicates that constraint term of the proposed algorithm can achieve a good agreement and cooperation with the basic objective function.

1. Introduction

When a dataset is provided with a few labeled samples in addition to a lot of unlabeled samples, it is necessary to develop effective approaches to model the relationship between two kinds. There are some potential approaches which can tackle such problems, e.g. semi-supervised clustering [1], transfer learning [2], kernel alignment measure [3], and transferred learning of the kernel [4], they were all constructed from the combination of labeled and unlabeled samples. The approach can effectively improve the quality of classification if the available a priori information is fully exploited. Semi-supervised clustering using a few labeled samples to guide the clustering process provides a novel enhanced concept beyond unsupervised and supervised clustering. In semi-supervised model-based classification, some semi-supervised clusterings take directly into account the available classification information [5], others utilize pairwise constraints between labeled samples [6–9]. The latter method introduces two types of constraints [10]: must-link and cannot-link constraints according to the two labeled samples must (or cannot) be assigned to one cluster or not. This provides the first basis for semi-supervised clustering of such kind, pairwise constraints then can be incorporated additively into unsupervised clustering algorithms as a penalty cost function of its objective function to obtain a new semi-supervised optimization problem. These constraints serve as supervised elements and provide general guidance for the clustering process towards more appropriate partitions.

In recent years, the method of pairwise constraints which has been proven to be an effective way to express the priori knowledge has attracted more and more interests. Pairwise constraints have been introduced into some basic algorithms and consequently obtained several different semi-supervised clustering algorithms. For example, there are semi-supervised hard clustering (PCKmeans [6], COP-Kmeans [7]), semi-supervised fuzzy clustering (AFCC [1]), and semi-supervised spectral clustering [11], etc. Specifically, AFCC suggested a semi-supervised fuzzy clustering by modifying the objective function of CA (clustering by competitive agglomeration) [12] with the available pairwise constraints. Unfortunately, its penalty cost function could not achieve an appropriate cooperation with the objective function of CA, consequently, AFCC was affected by the over adjustment of membership values leading to their deviation from the normal ranges. In order to overcome the above shortcoming, we propose a new semi-supervised fuzzy clustering algorithm (SCAPC) based on the redefined objective function. A new constraint penalty cost function is introduced to CA, as expected it can achieve a good agreement and cooperation with the basic objective function of CA.

* Corresponding author. Tel.: +86 510 85913612.
E-mail addresses: wu_xiaojun@yahoo.com.cn, cuifang_gao@163.com (X.-J. Wu).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.05.032

This paper is organized as follows: Section 2 briefly describes the unsupervised fuzzy clustering CA and the semi-supervised fuzzy clustering algorithm AFCC. Section 3 describes our new semi-supervised fuzzy clustering SCAPC in detail. Numerical tests and performance analysis are reported in Section 4. Finally, Section 5 makes the concluding remarks.

2. Related works

AFCC is stemmed from the clustering by competitive agglomeration (CA) [12]. Given a dataset $\Gamma = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \subset R^L$ which has N samples, each sample can be denoted as a vector with L attributes $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iL}]^T$. The objective function of CA is defined as follows:

$$J_{CA} = \sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) - \beta\sum_{k=1}^{C}\left[\sum_{i=1}^{N}u_{ik}\right]^2 \tag{1}$$

$$\text{Subject to the constraints} \quad \sum_{k=1}^{C}u_{ik} = 1 \quad i = 1, 2, \ldots, N \tag{2}$$

where C is the number of clusters, $\boldsymbol{\mu}_k = [\mu_{k1}, \mu_{k2} \ldots \mu_{kL}]^T$ is the kth cluster center, u_{ik} denotes the membership degree of the ith samples belonging to the kth cluster. Let $N_k = \sum_{i=1}^{N}u_{ik}$ be the cardinality of kth cluster expressed as the sum of membership degrees of the kth cluster. CA algorithm creates an environment in which clusters compete for samples based on cardinalities, and thus it controls the number of clusters. The choice of β is important in CA because it reflects the importance of the second term (competition term) relative to the first term (FCM term). Also because that the value of β should be chosen so that both terms are of the same magnitude order. In paper [12], β was chosen as the function of t:

$$\beta(t) = \frac{\eta_0 \exp(-t/\tau)}{\sum_{k=1}^{C}\left[\sum_{i=1}^{N}u_{ik}\right]^2}\left[\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\right] \tag{3}$$

where t is the iteration number. The definition of β in Eq. (3) is proportional to the ratio of the two terms, i.e.

$$\beta \propto \frac{\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, \mu_k)}{\sum_{k=1}^{C}\left[\sum_{i=1}^{N}u_{ik}\right]^2}$$

During the clustering process, β varies according to the iteration number t. The algorithm starts with a large value of β so that the competition term of the objective function dominates in the initial phase of training. The value of β decreases slowly to help CA to seek the optimal cluster number in the early iterations. When it is close to appropriate partitions, β becomes small and the first term is emphasized.

AFCC algorithm is an alternative semi-supervised fuzzy clustering which modifies the objective function of CA with the available pairwise constraints. According to the a priori information, the labeled samples are organized into two sets $\mathcal{M}$ and $\mathcal{C}$, the elements in $\mathcal{M}$ and $\mathcal{C}$ are pairs of data points. Let $\mathcal{M}$ be the set of must-link pairs such that $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{M}$ implies $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ should be assigned to the same cluster, and let $\mathcal{C}$ be the set of cannot-link pairs such that $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{C}$ implies $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ should be assigned to the different clusters. Note that the pairs in $\mathcal{M}$ and $\mathcal{C}$ are order-independent which means $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{M}$ is the same to $(\boldsymbol{x}_j, \boldsymbol{x}_i) \in \mathcal{M}$. The objective function of AFCC in paper [1] is given as:

$$J_{AFCC} = \sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \alpha\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{l=1,l\neq k}^{C}u_{ik}u_{jl} + \sum_{(x_i,x_j)\in\mathcal{C}}\sum_{k=1}^{C}u_{ik}u_{jk}\right) - \beta\sum_{k=1}^{C}\left[\sum_{i=1}^{N}u_{ik}\right]^2 \tag{4}$$

The optimization of J_{AFCC} is subject to the same constraints of Eq. (2). Objective function of AFCC is the combination of CA and the constraint function, the second term is just the penalty cost corresponding to the pairwise constraints. It calculates the violation cost of the constraints, including (1) penalty cost in $\mathcal{M}$ incurred by the must-link pairwise: the multiplier of corresponding membership values of two such points assigning to different clusters. (2) Penalty cost in $\mathcal{C}$ incurred by the cannot-link pairwise: the multiplier of corresponding membership values of two such points assigning to the same cluster.

In Eq. (4), parameter α is a weighted factor whose role is to maintain the relative importance of the supervision. Similarly, parameter β reflects the importance of the competition term. In paper [1] these two important parameters are given as:

$$\alpha = \frac{N}{M}\frac{\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k)}{\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2} \tag{5}$$

where M is the number of pairwise constraints, and β is chosen by

$$\beta(t) = \frac{\eta_0 \exp(-|t-t_0|/\tau)}{\sum_{k=1}^{C}\left[\sum_{i=1}^{N}u_{ik}\right]^2}\left[\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \alpha\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{l=1,l\neq k}^{C}u_{ik}u_{jl} + \sum_{(x_i,x_j)\in\mathcal{C}}\sum_{k=1}^{C}u_{ik}u_{jk}\right)\right] \tag{6}$$

Such expression of α using the normalized performance index (the sum of the squared distances between samples and centers divided by the sum of the squared memberships) intends to balance the contribution of supervised and unsupervised terms of J_{AFCC} [13]. The higher the normalized level is, the less important the supervision is. Unfortunately, this expression for α can not ensure the two terms are of the same magnitude order. So a potential drawback is that the membership values may not be confined to [0, 1] if they are over adjusted. This