Pattern Recognition 48 (2015) 953–967

Contents lists available at ScienceDirect

Pattern Recognition

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

Active constrained fuzzy clustering: A multiple kernels learning approach

Ahmad Ali Abin *, Hamid Beigy

Department of Computer Engineering, Sharif University of Technology, Azadi Ave., Tehran, Iran

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form
16 August 2014
Accepted 10 September 2014
Available online 20 September 2014

Keywords:
Constrained clustering
c-Means fuzzy clustering
Multiple kernels
Active constraint selection

ABSTRACT

In this paper, we address the problem of constrained clustering along with active selection of clustering constraints in a unified framework. To this aim, we extend the improved possibilistic c-Means algorithm (IPCM) with multiple kernels learning setting under supervision of side information. By incorporating multiple kernels, the limitation of improved possibilistic *c*-means to spherical clusters is addressed by mapping non-linear separable data to appropriate feature space. The proposed method is immune to inefficient kernels or irrelevant features by automatically adjusting the weight of kernels. Moreover, extending IPCM to incorporate constraints, its strong robustness and fast convergence properties are inherited by the proposed method. In order to avoid querying inefficient or redundant clustering constraints, an active query selection heuristic is embedded into the proposed method to query the most informative constraints. Experiments conducted on synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

1. Introduction

In recent years, constrained clustering has been emerged as an efficient approach for data clustering and learning the similarity measure between patterns [1]. It has become popular because it can take advantage of side information when it is available. Incorporating domain knowledge into the clustering by adding constraints enables users to specify desirable properties of the result and improves the robustness of clustering algorithm.

The first introduction of constrained clustering to the machine learning [2] focused on the use of instance-level pairwise constraints. Pairwise constraints specify whether two objects belong to the same cluster or not, known as the must-link (ML) constraints and the cannot-link (CL) constraints, respectively. Recent techniques in constrained clustering include integrating both clustering algorithm and learning the underlying similarity metric in a uniform framework [3], joint clustering and distance metric learning [4], topology preserving distances metric learning [5], Kernel approaches for metric learning [6], learning a margin-based clustering distortion measure using boosting [7], learning Mahalanobis distances metric [8,9], learning distances metric based on similarity information [10], and learning a distance metric transformation that is globally linear but locally non-linear [11], to mention a few.

Existing methods in constrained clustering reported the clustering performance averaged over multiple randomly generated constraints [2,5,7,8]. Random constraints do not always improve the quality of results [12]. In addition, averaging over several trials is not possible in many applications because of the nature of problem or the cost and difficulty of constraint acquisition. An alternative to get the most beneficial constraints for the least effort is to actively acquire them. There is a small range of studies on active selection of clustering constraints based on: "farthest-first" strategy [13], hierarchical clustering [14], theory of spectral decomposition [15], fuzzy clustering [16], Min–Max criterion [17], graph theory [18], and boundary information of data [19]. These methods choose constraints without considering how the underlying clustering algorithm utilizes the selected constraints. If we choose constraints independent of the clustering algorithm, it will have better performance for some algorithms but perform worse for some others.

This paper proposes a unified framework for constrained clustering and active selection of clustering constraints. It integrates the improved possibilistic *c*-Means (IPCM) [20] with a multiple kernels learning setting under supervision of side information. The proposed method attempts to address the limitation of IPCM to spherical clusters by incorporating multiple kernels. In addition, it immunizes itself to inefficient kernels or irrelevant features by automatically adjusting weight of kernels in an alternative optimization manner. In order to avoid querying inefficient or redundant constraints, an active query selection heuristic is embedded into the proposed method based on the measurement of Type-II mistake in clustering.

* Corresponding author. Tel.: +98 21 66166674.
E-mail addresses: abin@ce.sharif.edu (A.A. Abin), beigy@sharif.edu (H. Beigy).

http://dx.doi.org/10.1016/j.patcog.2014.09.008
0031-3203/

This heuristic attempts to query the most informative set of constraints based on the current state of the clustering algorithm. Altogether, the proposed method attempts to have the whole robustness against noise and outliers, immunization to inefficient kernels or irrelevant features, fast convergence rate, and selection of useful set of constraints in a unified framework. Experiments conducted on synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows: a brief overview of fuzzy clustering is provided in Section 2. We then introduce the proposed method in Section 3 and the proposed embedded active constraint selection heuristic is given in Section 4. Experimental results and time complexity analysis are presented in Section 5. Discussion on the proposed method is presented in Section 6. This paper concludes with conclusions and future works in Section 7.

2. Fuzzy clustering

Many clustering algorithms have been proposed over the past decades which perform hard clustering of data [21–24]. On the other hand, clusters may overlap in many real-world problems so that many items have the characteristics of several clusters [25]. In order to consider the overlaps among clusters, it is more natural to assign a set of memberships to each item, one for each cluster. This method is called fuzzy clustering. Fuzzy c-Means (FCM) algorithm is one of the most promising fuzzy clustering methods, which in most cases is more flexible than the corresponding hard clustering [26]. Given a dataset, $X=\{x_1,\dots,x_N\}$, where $x_i \in R^l$ and l is the dimension of feature vector, FCM partitions X into C fuzzy partitions by minimizing the following objective function:

$$J_{FCM}(U,V)=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^m d_{ci}^2 \tag{1}$$

where $V=(v_1,\dots,v_C)$ is a C-tuple of prototypes, d_{ci}^2 is the distance of feature vector x_i to prototype v_c, i.e. $\|x_i-v_c\|^2$, N is the total number of feature vectors, C is the number of partitions, u_{ci} is the fuzzy membership of x_i in partition c satisfying $\sum_{c=1}^{C}u_{ci}=1$, m is a quantity controlling the clustering fuzziness, and $U\equiv[u_{ci}]$ is a $C\times N$ matrix called fuzzy partition matrix, which satisfies three conditions $u_{ci}\in[0,1]$ for all i and c, $\sum_{i=1}^{N}u_{ci}>0$ for all c, and $\sum_{c=1}^{C}u_{ci}=1$ for all i. The original FCM uses the probabilistic constraint that the memberships of a data point across all partitions sum to one. While this is useful in creation of partitions, it makes FCM very sensitive to outliers or noise. Krishnapuram et al. proposed the possibilistic c-Means (PCM) clustering algorithm by relaxing the normalized constraint of FCM [27]. PCM minimizes the following objective function for clustering:

$$J_{PCM}(T,V)=\sum_{c=1}^{C}\sum_{i=1}^{N}t_{ci}^p d_{ci}^2+\sum_{c=1}^{C}\eta_c\sum_{i=1}^{N}(1-t_{ci})^p \tag{2}$$

where t_{ci} is the possibilistic membership of x_i in cluster c, $T\equiv[t_{ci}]$ is a $C\times N$ matrix called possibilistic partition matrix, which satisfies two conditions $t_{ci}\in[0,1]$ for all i and c and $\sum_{i=1}^{N}t_{ci}>0$ for all c, p is a weighting exponent for the possibilistic membership, and η_c are suitable positive numbers. The first term of $J_{PCM}(T,V)$ demands that the distances from data points to the prototypes be as low as possible, whereas the second term forces t_{ci} to be as large as possible. PCM determines a possibilistic partition, in which a possibilistic membership measures the absolute degree of typicality of a point in a cluster. PCM is robust to outliers or noise, because a far away noisy point would belong to the clusters with small possibilistic memberships, and consequently it cannot affect the resulting clusters significantly. However, its performance depends highly on a good initialization and has the undesirable tendency to produce coincident clusters. Zhang et al. proposed the improved possibilistic c-Means (IPCM) with strong robustness and fast convergence rate [20]. IPCM integrates FCM into PCM, so that the improved algorithm can determine proper clusters via the fuzzy approach while it can achieve robustness via the possibilistic approach. IPCM partitions X into C fuzzy partitions by minimizing the following objective function:

$$J_{IPCM}(T,U,V)=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^m t_{ci}^p d_{ci}^2+\sum_{c=1}^{C}\eta_c\sum_{i=1}^{N}u_{ci}^m(1-t_{ci})^p \tag{3}$$

Dealing with spherical clusters is the most important limitation of IPCM that can be eliminated by integration of IPCM with multiple kernels learning setting.

3. The proposed method

In the previous section, IPCM was explained as an efficient algorithm for data clustering dealing with partially overlapping clusters and noisy datasets. While IPCM is a popular soft clustering method, its effectiveness is largely limited to spherical clusters and cannot deal with complex data structures. Also, IPCM does not take side information into account to be used for constrained clustering. To provide a better clustering result, we propose a clustering algorithm that not only can deal with the linear inseparable and partially overlapping dataset, but also gets a better clustering accuracy under the noise interference. Also, the proposed algorithm is able to incorporate domain knowledge into the clustering algorithm to take advantage of the side information. To this aim, a new objective function is introduced to consider these issues.

To consider partially overlapped clusters and noise interference, the idea of IPCM is embedded into the new objective function. In order to take into account the side information, the objective function of IPCM is extended by two extra terms for the violation of the side information. Also, by applying multiple kernels setting, we attempt to address the problem of dealing with non-linear separable data, namely by mapping data with non-linear relationships to appropriate *feature spaces*. On the other hand, Kernel combination, or selection, is crucial for effective kernel clustering. Unfortunately, for most applications, it is not easy to find the right combination of the similarity kernels. Therefore, we propose a constrained multiple kernel improved possibilistic c-means (CMKIPCM) algorithm in which IPCM algorithm is extended to incorporate the side information with a multiple kernels learning setting. By incorporating multiple kernels and automatically adjusting the kernel weights, CMKIPCM is immune to ineffective kernels and irrelevant features. This makes the choice of kernels less crucial. Let $\psi(x)=\omega_1\psi_1(x)+\omega_2\psi_2(x)+\dots+\omega_M\psi_M(x)$ be a non-negative linear combination of M base kernels in kernel space Ψ to map data to an implicit feature space. The proposed method minimizes the following objective function for constrained clustering:

$$\begin{aligned}J_{\mathrm{CMKIPCM}}(\mathbf{w},T,U,V)=&\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^m t_{ci}^p(\psi(x_i)-v_c)^T(\psi(x_i)-v_c)\\&+\sum_{c=1}^{C}\eta_c\sum_{i=1}^{N}u_{ci}^m(1-t_{ci})^p\\&+\alpha\Bigg(\sum_{(i,j)\in\mathcal{M}}\sum_{c=1}^{C}\sum_{\substack{l=1\\l\neq c}}^{C}u_{ci}^m u_{lj}^m t_{ci}^p t_{lj}^p\\&+\sum_{(i,j)\in\mathcal{C}}\sum_{c=1}^{C}u_{ci}^m u_{cj}^m t_{ci}^p t_{cj}^p\Bigg)\end{aligned} \tag{4}$$

where $\mathcal{M}$ is the set of must-link constraints and $\mathcal{C}$ is the set of cannot-link constraints. $v_c\in\mathbb{R}^L$ is the center of c^{th} cluster in the implicit L-dimensional feature space and $V\equiv[v_c]_{L\times C}$ is a $L\times C$ matrix whose columns correspond to cluster centers. $\mathbf{w}=(\omega_1,\omega_2,\dots,\omega_M)^T$ is a vector consisting of kernel weights, which satisfies the condition $\sum_{k=1}^{M}\omega_k=1$. $U\equiv[u_{ci}]_{C\times N}$ is the fuzzy membership matrix whose