Pattern Recognition 47 (2014) 1443–1458

Contents lists available at ScienceDirect

Pattern Recognition

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

Active selection of clustering constraints: a sequential approach

Ahmad Ali Abin *, Hamid Beigy

Department of Computer Engineering, Sharif University of Technology, Azadi Avenue, Tehran, Iran

ARTICLE INFO

Article history:
Received 20 January 2013
Received in revised form
21 August 2013
Accepted 30 September 2013
Available online 9 October 2013

Keywords:
Active constraint selection
Constrained clustering
Pairwise constraints
Data description

ABSTRACT

This paper examines active selection of clustering constraints, which has become a topic of significant interest in constrained clustering. Active selection of clustering constraints, which is known as minimizing the cost of acquiring constraints, also includes quantifying utility of a given constraint set. A sequential method is proposed in this paper to select the most beneficial set of constraints actively. The proposed method uses information of boundary points and transition regions extracted by data description methods to introduce a utility measure for constraints. Since previously selected constraints affect the utility of remaining candidate constraints, a method is proposed to update the utility of remaining constraints after selection of each constraint. Experiments carried out on synthetic and real datasets show that the proposed method improves the accuracy of clustering while reducing human interaction.

1. Introduction

Data clustering is an exploratory and descriptive data analysis technique with a long history in a variety of scientific fields [1]. It is presented with data instances that must be grouped according to a notion of similarity [2]. Clustering is fundamentally done with making some initial assumptions on distance metric, data structure, number of clusters, data distribution, and so on. If there is no correspondence between these assumptions and the actual model of clusters, the algorithm results in poor clusters. Recently, constrained clustering has become popular because it can take advantage of side information when available. Incorporating domain knowledge into the clustering by addition of constraints enables users to specify desirable properties of the result and also improves the robustness of the clustering algorithm.

There are several families of constraints but the instance-level constraints are the most used. Wagstaff et al. [2] introduced side information in two types of instance-level constraints: must-link (ML) and cannot-link (CL) constraints. A must-link (positive) constraint requires two objects to be grouped into the same cluster while a cannot-link (negative) constraint requires two objects to be put in different clusters. The inclusion of instance-level constraints allows the user to precisely state which objects should belong to the same cluster and which objects should not, without the need to explicitly state what these clusters are. In addition, the side information may occur at different levels such as class labels for a subset of objects, knowledge about clusters' position, clusters' identity, minimum or maximum size of clusters, and distribution of data [3].

Existing methods, which use constraints in the form of must-link and cannot-link constraints, can be grouped into two main categories: constraint-based and distance-based methods. In the first category [2,4], the constraints state whether two instances should be grouped into the same cluster or not, and the clustering algorithm is adapted so that the available constraints bias the search for a suitable clustering of data. Algorithms in the second category are initially trained to learn a proper distance measure satisfying the given constraints and then use the learnt measure for clustering of data. Recent techniques in this category include: joint clustering and distance metric learning [5], topology preserving distances metric learning [6], kernel approaches for metric learning [7], learning a margin-based clustering distortion measure using boosting [8], learning Mahalanobis distances metric [9,10], learning distances metric based on similarity information [11], and learning a distance metric transformation that is globally linear but locally non-linear [12], to mention a few. Use of pairwise constraints is not limited only to clustering applications. Several authors used pairwise constraints in semi-supervised feature selection [13], and dimensionally reduction [14], to mention a few.

While there is a large body of researches on constrained clustering algorithms [2,4–9,11,12,15,16], recently some more fundamental issues have emerged [17]. Three important issues in constrained clustering are: (1) quantifying the utility of a given constraint set, (2) minimizing the cost of constraint acquisition, and (3) propagating the constraint information to nearby regions in order to reduce the number of needed constraints [17]. The second issue comprises the first one and refers to minimizing the

* Corresponding author. Tel.: +98 21 66166698.
E-mail addresses: abin@ce.sharif.edu, ali_abin@yahoo.com (A.A. Abin), beigy@sharif.edu (H. Beigy).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.09.034

cost of constraint acquisition. Existing methods in constrained clustering assumed that the algorithm is fed with a suitable passively chosen set of constraints [2,6,8,9]. They reported the clustering performance averaged over multiple randomly generated constraints. This is not always an applicable assumption. Randomly selected constraints do not always raise the quality of clustering results [18]. In addition, averaging over several trials is impossible in many problems because of the nature of the given problems or the cost of constraint acquisition. On the other hand, there are $\frac{1}{2}N(N-1)$ possible constraints on a dataset with N instances, and constraint specification can be burdensome for large datasets. An alternative way to easily find the most beneficial constraints is to actively acquire them. There are small range of studies on active selection of clustering constraints, which include active selection of constraints based on: "farthest-first" strategy [19], hierarchical clustering [20], theory of spectral decomposition [21], fuzzy clustering [22], Min–Max criterion [23], and graph theory [24]. The performance of each of them highly depends on the underlying assumptions like the data structure, the distance metric and so on. These methods ignore the effect of previously chosen constraints on the utility of remaining candidate constraints but this paper directly addresses the constraint utility dependencies as an important issue in constrained clustering.

This paper proposes a method to improve active selection of clustering constraints. Our proposed method is based on the sequential selection of constraints such that the selection heuristic takes into account the already chosen constraints. The information of boundary points and transition regions of data is used to introduce a time-varying utility measure for constraints. The efficiency of the selected constraints is evaluated in conjunction with some constrained clustering algorithms on some artificial and real datasets. Experimental results show the superiority of the proposed method on the chosen constraints, which increase the accuracy of methods in constrained clustering.

The rest of this paper is organized as follows. The related work is discussed in Section 2. The proposed constraint selection method is given in Section 3. Experimental results are presented in Section 4. This paper concludes with conclusions and future works in Section 5.

2. Related work

Active learning has a long history in supervised learning algorithms. The key idea behind active learning is that a machine learning algorithm can perform better with less training if it is allowed to query the labels of some instances. In the statistics literature active learning is sometimes also called optimal experimental design [25]. Recently, active learning approaches are used in constrained clustering problem. Few studies reported the result of using active learning for constrained clustering [19–24,26]. Klein et al. [20] suggested an active constraint selection method in which a hierarchical clustering algorithm identifies the m best queries that should be asked from an expert.

Basu et al. [19] proposed an algorithm for active selection of constraints using the farthest first query selection (FFQS) algorithm. FFQS has two phases: *Explore* and *Consolidate*. Let K be the true number of clusters in a dataset. The exploration phase explores the given data to find K pairwise disjointed non-null neighborhoods (as skeleton of the clusters), belonging to different clusters. A preference to cannot-link queries is given by choosing the farthest point from the existing disjointed neighborhoods. The exploration will continue until K points are found such that there is a cannot-link between each pair of these K points. This phase is then followed by the consolidate phase. The consolidate phase selects non-skeleton points randomly and queries them against each point in the skeleton, until a must-link query is obtained.

Mallapragada et al. [23] generalized FFQS by introducing Min–Max criterion. Their method (referred to as MMFFQS) altered the consolidate phase of FFQS. The exploration in MMFFQS is done in the same way as FFQS but the random point selection of the consolidation phase is replaced with selection of data point with maximum uncertainty in cluster assignment. Both FFQS and MMFFQS have problem with unbalanced datasets or datasets containing a large number of clusters [24].

ACCESS [21] is an active constrained clustering technique which examines the eigenvectors derived from similarity matrix of data. ACCESS uses the theory of spectral decomposition to identify data items that are likely to be located on boundaries of clusters, and for which providing constraints can resolve ambiguity in the clustering. ACCESS identifies two types of informative points: (1) sparse points and (2) close and distant boundary points. The sparse points are identified by evaluating the first m eigenvectors of the similarity matrix (where m depends on how many sparse sub-clusters are found in the dataset), and the close and distant boundary points are identified by evaluating the $(m+1)$th eigenvector of the similarity matrix. These informative points are then used by ACCESS for active selection of constraints. However, limitation of ACCESS on problems with two clusters can be mentioned as an important shortcoming.

AFCC [22] as another active constrained clustering method minimizes a competitive cost function with fuzzy terms corresponding to pairwise constraints. AFCC uses an active method based on the least well-defined cluster to find the most informative must-link or cannot-link constraints. Fuzzy hyper-volume measure is used by AFCC to identify the least well-defined cluster and objects located on frontier of this cluster. For each object lying on the frontier, the closest cluster corresponding to its second highest membership value is found and the user is then asked whether one of these objects should be (or not) in the same cluster as the nearest object from the closest cluster. AFCC will fail when there are clusters with complex structure (shapes, distributions) in data.

Vu et al. [24,26] proposed an active query selection method based on the ability to separate between clusters. We refer to this algorithm as ASC. ASC has three basic steps: (1) the best candidate query in sparse regions of the dataset is determined by a k-nearest neighbor graph, (2) a constraint utility function is used in a query selection process, and (3) a propagation procedure is used to propagate each query to generate several constraints and limit the number of candidate constraints. The propagation procedure discovers new constraints from the information stored in already chosen constraints using the notion of strong paths. Subsequently, the size of the candidate set is reduced by a refinement procedure that removes constraints between objects that are likely to be in the same cluster. Specifically, the refinement procedure removes candidate constraints that are linked by a strong path.

The above-mentioned methods form a range of studies performed in active selection of clustering constraints. Each method considered a basic assumption on utility of constraints. Their applicability on a specific problem is highly dependent on correlation between their assumption and the actual structure of data.

3. The proposed active constraint selection method

In the literature, several methods have been proposed for active constraint selection but they ignore the constraint utility dependencies. In this section, we propose a sequential approach for active constraint selection (SACS) that considers dependencies among the constraints. The proposed approach is based on the following two assumptions.