Neurocomputing 000 (2017) 1–24

Contents lists available at ScienceDirect

Neurocomputing

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

Robust semi-supervised clustering with polyhedral and circular uncertainty

Derya Dinler*, Mustafa Kemal Tural

Industrial Engineering Department, Middle East Technical University, Dumlupinar Bulvari, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 28 February 2016
Revised 26 April 2017
Accepted 27 April 2017
Available online xxx

Keywords:
Clustering
Uncertainty
Semi-supervised learning
Second order cone programing
Heuristics

ABSTRACT

We consider a semi-supervised clustering problem where the locations of the data objects are subject to uncertainty. Each uncertainty set is assumed to be either a closed convex bounded polyhedron or a closed disk. The final clustering is expected to be in accordance with a given number of instance level constraints. The objective function considered minimizes the total of the sum of the violation costs of the unsatisfied instance level constraints and a weighted sum of squared maximum Euclidean distances between the locations of the data objects and the centroids of the clusters they are assigned to. Given a cluster, we first consider the problem of computing its centroid, namely the centroid computation problem under uncertainty (CCPU). We show that the CCPU can be modeled as a second order cone programing problem and hence can be efficiently solved to optimality. As the CCPU is one of the key ingredients of the several algorithms considered in this paper, a subgradient algorithm is also adopted for its faster solution. We then propose a mixed-integer second order cone programing formulation for the considered clustering problem which is only able to solve small-size instances to optimality. For larger instances, approaches from the semi-supervised clustering literature are modified and compared in terms of computational time and quality.

1. Introduction

Clustering is an unsupervised data mining technique which is used to partition unlabeled data objects into groups (clusters) according to their similarity. The aim of a clustering algorithm is to obtain "natural" structures inherent in the data. As clustering algorithms are completely unguided, the structures obtained are not always relevant or useful to the user.

In some cases, the user (or the data analyst) may have some a priori knowledge about the underlying structure of the data in the form of *must-link* and/or *cannot-link constraints*. These constraints are called as *instance level constraints* and can be used to guide the clustering algorithm in order to obtain more useful or relevant clustering of the data. A must-link constraint between two data objects is used to indicate that the objects are to be placed in the same cluster. A cannot-link constraint between two data objects, on the other hand, is used to indicate that the objects are to be placed in different clusters. Clustering algorithms that use such limited supervision to improve the clustering results are called as *semi-supervised clustering* algorithms.

A common approach in clustering is to consider an optimization problem that minimizes an objective function which is a function of *within cluster similarity* and/or *between cluster dissimilarity* and solve the resulting problem using exact algorithms, heuristics, or approximation algorithms. Different objective functions have been employed in the literature. The most commonly used one is to minimize the sum-of-squared (Euclidean) distances between each data object and the *centroid* of the cluster the object belongs to. Here, the centroid of a cluster can be considered as the representative point of the cluster and its coordinates, in the case with (fixed) point data objects, are obtained by averaging the coordinates of the data objects in the cluster (this ensures that the sum-of-squares objective function is minimized).

Traditional clustering algorithms use only (fixed) data points as input data objects. In this paper, we assume that there is uncertainty in the location of a data object and the uncertainty set is either a closed convex bounded polyhedron, i.e., a polytope, or a closed disk. So, the data objects considered are not points but rather regions in the Euclidean space. In particular, the problem considered in this paper is to partition a given number of polytopes and disks in the Euclidean space into a given number of clusters and find the cluster centroids so as to minimize the sum-of-squares objective function in the presence of instance level constraints.

* Corresponding author.
E-mail addresses: dinler@metu.edu.tr (D. Dinler), tural@metu.edu.tr (M.K. Tural).

http://dx.doi.org/10.1016/j.neucom.2017.04.073
0925-2312/

Please cite this article as: D. Dinler, M.K. Tural, Robust semi-supervised clustering with polyhedral and circular uncertainty, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.04.073

Consider a clustering algorithm that partitions a given number of regional data objects into a given number of clusters. The distance between an uncertain data object and the representative centroid is subject to change for each realization of the location of the data object. In this paper, our aim is to minimize the sum-of squares objective function in the worst-case. In other words, for any realization of the locations of the data objects, the evaluated objective function should not be greater than the optimal objective function value. For this reason, the distance between the centroid of a cluster and a regional data object is measured in terms of the maximum distance between them. The objective function used in this paper is also meaningful in the sense that any realization of a data object is not expected to be very far away from the representative cluster centroid as the worst-case scenario is considered. We use the term *robust* to emphasize the fact that worst-case scenarios are accounted for. Note that, for the problem considered, the solutions obtained will be sensitive to the outliers or extreme observations as in the case with the classical solution approaches for the minimum sum-of-squares clustering problem with data points. We assume in this paper that the regional data objects are noise free and do not consider handling robustness with respect to small changes in the locations of the data objects.

A related problem that is not studied in this paper is the problem of minimizing the maximum of the maximum distances between the regional data objects and the cluster centers (the coordinates of a center of a cluster are not necessarily the average of the coordinates of the data objects in the cluster even when each data object is a fixed point). This problem is a generalization of the well studied clustering/facility location problem that minimizes the maximum radius among the clusters which is known as the k-center problem (k represents the number of clusters to be formed), see e.g., [16]. The problem studied in this paper, on the other hand, reduces to the minimum-sum-of-squares (semi-supervised) clustering problem when all the regional data objects are points in the Euclidean space.

Clustering of data objects whose locations are uncertain (represented by probability distributions) is an important issue and appears in several real life settings. Consider, for example, clustering of digital cameras whose different aspects (e.g., image quality, battery performance) are rated by users [45]. The ratings of each camera can be represented by a multivariate probability distribution on a multidimensional region. If the worst-case scenarios are important, then the density of the distribution is of no relevance, but rather the multidimensional region that the distribution is defined over (where the density is nonzero) is important. If certain cameras are (not) to be placed in the same cluster, some instance level constraints can be introduced to guide the clustering algorithm in the direction of the desired clustering.

As a second example, consider forests in a certain region that are to be clustered to build a given number of fire stations. In this case, each forest can be thought of as a region and the location of a possible fire can be represented by a probability distribution over the forest. In this case, if clustering is performed using expected distances between forests and fire stations, then the point of a fire may turn out to be very far away from the station. In cases where the worst-case scenarios are important, robust solutions would be preferred over solutions obtained using expected distances.

Another example would be clustering of cities based on atmospheric conditions like temperature and humidity on a particular month. As the atmospheric conditions may vary during a specific month, each city may be represented by a multivariate probability distribution on a multidimensional region which is not necessarily a hyper-rectangle as temperature and humidity are not independent.

The contributions of this paper can be summarized as follows.

- A new robust clustering problem is introduced that aims to cluster regional data objects in the presence of instance level constraints and the possible application areas are summarized.
- The centroid computation problem is investigated: a second order cone programing formulation which is an extension of a previous formulation and a novel subgradient method are proposed for its solution. The subgradient method is fine-tuned and these two solution methods are computationally compared on several instances.
- For the considered semi-supervised clustering problem, a novel mixed-integer second order cone programing formulation is proposed. The proposed formulation is shown to solve small size instances (with 10 and 20 rectangular data objects) of the problem within about a minute.
- Six different semi-supervised clustering algorithms from the literature (five of them are k-means based and one of them is agglomerative hierarchical clustering based) are modified to make them applicable for the case with regional data objects. The subgradient method is utilized in these algorithms for centroid computations.
- For k-means based algorithms, three different initialization procedures are utilized and are compared on several instances.
- Four different instance level constraint generation techniques are used and for different algorithms the combination of initialization procedure and instance level generation technique resulting in the best performance is computationally experimented.
- Different numbers of instance level constraints are provided to the algorithms to see their effects on the performance measures.
- The algorithms are compared utilizing six different performance measures on artificial and real-life datasets.

The rest of the paper is organized as follows: in Section 2, we provide a literature review on semi-supervised clustering and clustering/facility location problems/algorithms that deal with regional data objects. In Section 3, we introduce the notation used throughout the paper. In Section 4, we discuss solution approaches that compute the centroid of a given cluster consisting of regional data objects. In Section 5, we first model the considered clustering problem as a mixed integer second order cone programing problem which is only able to solve small size instances. After reviewing some solution approaches from the semi-supervised clustering literature proposed mainly for point data objects, we then discuss modified versions of them that can handle regional data objects. The computational studies are presented in Section 6 and we conclude in Section 7 with some future research directions.

2. Literature review

Most clustering algorithms can be categorized as hierarchical or partitional. Hierarchical clustering algorithms build a hierarchy of clusters by merging (agglomerative methods) or splitting (divisive methods) clusters successively [32]. The hierarchy of clusters is usually represented by a tree known as dendrogram. By cutting the dendrogram at the proper level, clustering with the desired number of clusters can be obtained. For more on hierarchical clustering algorithms, the reader is referred to [55,58,75]. Partitional clustering algorithms, on the other hand, attempts to construct a one-level clustering of the data objects without a hierarchy.

The problem of clustering a given number of point data objects so that the sum-of-squares objective function is minimized is NP-hard in general [56] and thereof heuristic solution approaches have been widely used. The k-means algorithm [54], which is among the most popular data mining algorithms, and its variants are the most commonly used partitional clustering heuristics proposed for

Please cite this article as: D. Dinler, M.K. Tural, Robust semi-supervised clustering with polyhedral and circular uncertainty, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.04.073