Information Sciences 339 (2016) 353–368

Contents lists available at ScienceDirect

Information Sciences

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

A Rough–Fuzzy approach for Support Vector Clustering

Ramiro Saltos*, Richard Weber

Department of Industrial Engineering, FCFM, Universidad de Chile, República 701, Santiago de Chile, Chile

ARTICLE INFO

Article history:
Received 13 November 2014
Revised 17 December 2015
Accepted 25 December 2015
Available online 8 January 2016

Keywords:
Fuzzy sets
Rough sets
Support Vector Clustering
Data mining

ABSTRACT

Support Vector Clustering (SVC) is an important density-based clustering algorithm which can be applied in many real world applications given its ability to handle arbitrary cluster silhouettes and detect the number of classes without any prior knowledge. However, if outliers are present in the data, the algorithm leaves them unclassified, assigning a zero membership degree which leads to all these objects being treated in the same way, thus losing important information about the data set. In order to overcome these limitations, we present a novel extension of this clustering algorithm, called Rough–Fuzzy Support Vector Clustering (RFSVC), that obtains rough–fuzzy clusters using the support vectors as cluster representatives. The cluster structure is characterized by two main components: a lower approximation, and a fuzzy boundary. The membership degrees of the elements in the fuzzy boundary are calculated based on their closeness to the support vectors that represent a specific cluster, while the lower approximation is built by the data points which lie inside the hyper-sphere obtained in the training phase of the SVC algorithm. Our computational experiments verify the strength of the proposed approach compared to alternative soft clustering techniques, showing its potential for detecting outliers and computing membership degrees for clusters with any silhouette.

1. Introduction

Clustering is one of the most important data mining tasks. Its objective is to find natural groups in a given data set in which the observations would be homogeneous within each group and heterogeneous between groups. Many clustering algorithms have been proposed in the literature [10,14,27,34,36,37], which can be grouped into two categories: hard clustering, and soft clustering. Their main difference is that in hard clustering an object has to belong exactly to one cluster while this constraint is relaxed in soft approaches. In certain applications, the hard clustering approach may not be adequate given the nature of the problem. Consequently, soft clustering could grant more flexibility in deriving adequate solutions.

Since their introduction, fuzzy sets [40] and rough sets [26] have shown their particular advantages when ambiguity and uncertainty have to be dealt with [27]. Fuzzy C-Means [5] and Rough C-Means [20] are very common representatives of soft clustering algorithms, and their derivatives have been applied in many areas. However, their use is still limited by some characteristics, such as clusters with spherical shapes, the fact that the sum of the membership values of an object has to be equal to 1 (Fuzzy C-Means), the need to know the number of clusters beforehand, and that the data points identified as outliers are not classified accordingly.

* Corresponding author. Tel.: +56 9 54362841.
E-mail addresses: ramiro.saltos@ing.uchile.cl, rjsaltos1989@gmail.com (R. Saltos), rweber@dii.uchile.cl (R. Weber).

http://dx.doi.org/10.1016/j.ins.2015.12.035
0020-0255/© 2016 Elsevier Inc. All rights reserved.

On the other hand, Support Vector Clustering [4], an important density-based clustering algorithm, uses the support vector machines philosophy to find the clusters' cores, gaining the capability of detecting classes of any shape without having to know in advance the number of clusters. However, this algorithm treats all data points that do not belong to one of the clusters found in the same way, indicating them just as outliers. Especially, in the presence of scattered data this turns out to be a major drawback since important information could be lost by not differentiating among outliers.

To overcome these deficiencies, we propose a novel clustering algorithm called Rough–Fuzzy Support Vector Clustering, a generalization of Support Vector Clustering, that as will be shown, has the following advantages: similar to traditional SVC, any cluster shape can be detected and it is not necessary to know the number of clusters in advance, since the basic idea relies on the concept of density of data points. Additionally, those data points that are not clearly assigned to one of the clusters (here considered as outliers) get membership values to all clusters according to their distances in the higher-dimensional feature space. These membership values provide important information about the outliers, which in many applications could be the most critical cases.

The remainder of this paper is arranged as follows: Section 2 presents the traditional Support Vector Clustering algorithm and provides an overview of the state-of-the-art of its soft computing variations. Section 3 introduces the proposed method called Rough–Fuzzy Support Vector Clustering and explains its basic ideas in detail. Experimental results using RFSVC, and alternative methods, are presented in Section 4. Finally, Section 5 contains a summary of this paper, provides its main conclusions, and indicates future developments.

2. Literature overview on Support Vector Clustering

In this section, we present a general introduction to Support Vector Clustering. Then, we provide the respective algorithm's mathematical description in order to have the basis for developing our rough–fuzzy clustering method. Finally, we comment on recent studies related to our approach to emphasize its importance.

2.1. General introduction to Support Vector Clustering

Let $X = \{\mathbf{x}_i \in \mathcal{R}^d / i = 1, 2, \ldots, N\}$ be the set of N data points and $\mathcal{R}^d$ be the data space. The traditional Support Vector Clustering algorithm groups the elements in set X into clusters, interpreting the solution of the Support Vector Domain Description [33] as cluster cores, and assigns each individual point to its nearest core to generate the final clusters [9]. This is achieved using a two-phase algorithm consisting of a *training phase* and a *labeling phase*.

During the training phase, following the ideas proposed by Tax and Duin [33], data points are projected from the original data space to some higher-dimensional space looking for the hyper-sphere with a minimal radius that encloses most of the data points, as shown in Fig. 1(a). When the enclosing sphere is found, three kinds of data points can be identified: support vectors (SV), bounded support vectors (BSV), and inside data points (ID). Support vectors are data points whose images lie on the surface of the enclosing sphere, while bounded support vectors lie outside the hyper-sphere, and inside data points belong to its interior.

After that, the images of data points are projected back from the higher-dimensional space to the original data space where support vectors now define a set of contours that enclose data points. This completes the training phase (Fig. 1(b). Finally, the labeling phase identifies the different clusters found during training and allows building a {0, 1}-membership matrix which indicates to which cluster each data point belongs.

2.2. Mathematical description of Support Vector Clustering

In this section, following Ben-Hur's work, we present the mathematical description of the training phase and the labeling phase of Support Vector Clustering algorithm. For more details and proofs, see [4].

2.2.1. Training phase

In the training phase, a quadratic optimization problem is solved in order to find the hyper-sphere with minimal radius in a higher-dimensional space that encloses the images of the available data points from the original space. The model is formulated as follows:

$$Min\ R^2 + C\sum_{i=1}^{N}\xi_i \tag{1}$$

$$s.t.\ \ \| \phi(\mathbf{x}_i) - \boldsymbol{a} \|^2 \le R^2 + \xi_i \quad \forall i = 1, \ldots, N \tag{2}$$

$$\xi_i \ge 0 \quad \forall i = 1, \ldots, N \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{a}$ is the center of the hyper-sphere, ϕ is the non-linear function that projects data from the original space to the higher-dimensional space, ξ_i are slack variables that relax the constraints to allow some data points to lie outside the sphere, R is the sphere's radius, and $C \in [0, 1]$ is a constant penalty parameter.