Use the Core Clusters for the Initialization of the Clustering Based on One-Class Support Vector Machine

Lei Gu¹,²

¹ The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 200092, China
² School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China
gulei@njupt.edu.cn

Abstract. The clustering method based on one-class support vector machine has been presented recently. Although this approach can improve the clustering accuracies, it often gains the unstable clustering results because some random datasets are employed for its initialization. In this paper, a novel initialization method based on the core clusters is used for the clustering algorithm based one-class support vector machine. The core clusters are gained by constructing the neighborhood graph and they are regarded as the initial datasets of the clustering algorithm based one-class support vector machine. To investigate the effectiveness of the proposed approach, several experiments are done on four datasets. Experimental results show that the new presented method can improve the clustering performance compared to the previous clustering algorithm based on one-class support vector machine and k-means approach.

Keywords: Initialization, One-class support vector machine, Neighbor graph, Core clusters, Clustering methods.

1 Introduction

The aim of data clustering methods is to divide data into several homogeneous groups called clusters, within each of which the similarity or dissimilarity between data is larger or less than data belonging to different groups[1]. Unsupervised clustering partitions all unlabeled data into a certain number of groups on the basis of one chosen similarity or dissimilarity measure[2,3]. Different measure of the similarity or dissimilarity can lead to various clustering methods such as k-means[4], fuzzy c-means[5], mountain clustering, subtractive clustering[6] and neural gas[7]. In these traditional clustering algorithms, k-means, which can be easily implemented, is the best-known squared error- based clustering algorithm. Recently, a novel kernel method for clustering based on one-class support vector machine(COSVM) was presented in [8]. This kernel-based clustering method can be implemented in a similar way to the classical k-means and use a one-class support vector machine as the
Use the Core Clusters for the Initialization of the Clustering

description of each cluster rather than the center of several data. Experiments on real
datasets show that the clustering algorithm in [8] is valid and can have encouraging
performance.

However, this COSVM clustering algorithm is easily affected by the random initial
datasets. So it often has the unstable clustering performance. To solve this problem,
an initialization method based on the core clusters is applied to the COSVM
(CC-COSVM) in this paper. Several core clusters can be produced by constructing the
$\sigma$-neighborhood graph by the way in [9] and they also are regarded as the initial
datasets of the CC-COSVM. Data points included by the core clusters are a part of the
whole dataset. The number of core clusters is equal to the number of clusters.
Experimental results show that the proposed CC-COSVM approach can not only
obtain the stable clustering performance, but also improve the clustering accuracies
when compared to the COSVM and KM method.

The remainder of this paper is organized as follows. Section 2 reports the COSVM
clustering algorithm. In Section 3, the new proposed CC-COSVM clustering approach
is formulated. Some experimental results are shown in Section 4 and Section 5 give
some conclusions.

2 The COSVM Clustering Algorithm

The COSVM clustering algorithm is a kernel-based clustering approach. Inspired by
the KM method, it can gain the better clustering accuracies than the KM. The key step
is that the one-class support machine is trained.

Firstly, the one-class support vector machine method is introduced. One-class
support vector machine is a kernel-based data domain description method. It tries to
find the smallest sphere containing all input data in the feature space. Assume that a
nonempty set of cluster $m$ in the $d$-dimensional space $\mathbb{R}^d$ is $X^m$ and
$X^m = \{x_1^m, x_2^m, \cdots, x_n^m\}$. Now we construct a smallest sphere $S_m$ for cluster $m$
that can enclose all points $x_i^m$ ($x_i^m \in X^m, i = 1, 2, \cdots, n$). This problem is
considered as a quadratic optimization as follows[10]:

$$\min_{a, R} R \text{ s.t. } (x_i^m - a)^T (x_i^m - a) \leq R^2, i = 1, 2, \cdots, n$$

(1)

where $a$ is the center of the sphere $S_m$ and $R$ is its radius. The constraints
introduce slack variables $\xi_i$ as follows:

$$\min_{a, R, \xi} R + C \sum_{i=1}^{n} \xi_i \text{ s.t. } (x_i^m - a)^T (x_i^m - a) \leq R^2 + \xi_i$$

(2)

where $\forall i, \xi_i \geq 0$ and the variable $C$ gives the trade-off between the volume of the
sphere and the number of target objects rejected. This is a convex optimization