Optik 126 (2015) 4964–4970

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Automated clustering by support vector machines with a local-search strategy and its application to image segmentation

Chih-Hung Wu, Chih-Chin Lai*, Chun-Yen Chen, Yan-He Chen

Department of Electrical Engineering, National University of Kaohsiung, 700, Kaohsiung University Road, Nan-Tzu District, Kaohsiung 811, Taiwan

ARTICLE INFO

Article history:
Received 17 October 2014
Accepted 9 September 2015

Keywords:
Automated clustering
Support vector machine
Cluster validity index
Local-search strategy
Image segmentation

ABSTRACT

Deciding a rational number of clusters for the problems to be solved and determining the parameters associated with the clustering algorithms are two critical issues in the configuration of data clustering. Usually, a manual trial-and-error manner is used to induce a feasible configuration. This paper presents an automated clustering method, which determines the clustering configuration automatically. The proposed method is based on the techniques of support-vector machines with a local-search strategy. It starts with running one-class support-vector machines (OCSVM) to partition input data into a random number of clusters. When a result is obtained, the "local-search" mechanism launches several rounds of OCSVM each of which works with a new clustering configuration. Each new configuration is from the current configuration with incrementally modifications. The clustering results obtained from the local searches are post-evaluated by specific clustering validity index and the best one is retained. The clustering configuration of the best result is used by OCSVM for the clustering afterwards. Such a clustering process iterates until no better result can be obtained. This paper describes the clustering algorithm and compares three clustering validity indices, i.e. distance-based index, Davies–Bouldin index, and Xie–Beni index, on their effectiveness. The performance of the proposed method is demonstrated on the segmentation of several aerial images. Experimental results show that the proposed approach is feasible and effective for image segmentation.

1. Introduction

Cluster analysis has been effectively used in many applications such as pattern recognition, information retrieval, data mining, etc. Whatever the clustering methods used, two issues are critical: (1) how efficiently and reasonably the objects or data are processed by the clustering algorithm, and (2) how the number of clusters is determined. A clustering algorithm partitions input data into a number of clusters. It is usually assumed that the cluster number is known *a priori*. Too many or too few clusters may be less informative. Besides, a clustering algorithm usually works with a set of hyper-parameters that guides the algorithm how to search the solution space. Setting a set of effective parameters also dominates the performance of the algorithm. To obtain a good clustering result that is reasonable and helpful for advanced analysis, the "configuration" of clustering, i.e. deciding the cluster number and parameter settings, should be carefully considered. However, it is difficult for users to have exact ideas about what the best clustering configuration is in real-life applications. Hence, most users are forced to use a trial-and-error means to find feasible configurations for clustering. Such configurations may need to be revised frequently if clustering results are not good enough. Therefore, designing effective clustering methods that can automatically determine a good clustering configuration presents real challenges.

This paper presents a clustering method based on the techniques of support-vector machines (SVMs) [1] which decides the clustering configuration automatically using the local-search strategy. SVM has been emerging as a popular classifier due to its efficiency and ability to handle complex problems. A simple one-class support-vector machine (OCSVM), which is a version of SVM, is employed as the core clustering algorithm. The clustering process starts with a round of clustering by OCSVM which partitions data into a random number of clusters. Then, the local-search strategy is applied as follows. A tracking mechanism launches several rounds of OCSVM clustering, each of which works with a new set of configuration. These new configurations are copies of the current configuration with incremental modifications. The clustering results by these new configurations are evaluated and compared, and the best one (local-optimal) is selected as the new state. Based on the new state, new configurations are selected for clustering

* Corresponding author. Tel.: +886 75919771.
E-mail addresses: johnw@nuk.edu.tw (C.-H. Wu), cclai@nuk.edu.tw (C.-C. Lai).

http://dx.doi.org/10.1016/j.ijleo.2015.09.143
0030-4026/

afterwards and the local-search strategy iterates until no better result can be found.

SVM transforms data points from data space to a high dimensional feature space so that the data points in the feature space become linearly separable for partitioning. To better evaluate the clustering results in the original data space, the clustering results of OCSVM are post-evaluated by specific clustering validity index. Three clustering validity indices, i.e. distance-based index, Davies–Bouldin index, and Xie–Beni index, are incorporated with OCSVM. The feasibility of the proposed approach is demonstrated on the segmentation of several real aerial images. Comparisons on the effectiveness of these indices and various parameters settings on the proposed methods are conducted. Experimental results demonstrate the feasibility of the proposed approach.

2. Related work

This section presents several clustering methods based on soft-computing and machine learning for image clustering. Awad et al. present [2] a hybrid clustering method which combines genetic algorithm (GA) and fuzzy C-means (FCM). The GA part is enhanced using Hill-climbing, randomizing, and modified mutation operators for deciding the number of clusters. The FCM is modified for fast deciding the cluster centres. Experiments on several satellite images reveal that FCM-GA segmentation method gives robust and reliable results, and is more time efficient. In [3], a multi-component image segmentation method is developed using a nonparametric unsupervised self-organizing map (SOM) and hybrid genetic algorithm (HGA). SOM is used for feature extraction; while HGA is used for automated clustering without *a priori* knowledge. Clustering-based image segmentation by considering membership-connectedness (MC) with a fuzzy clustering stage is presented in [4]. The segmentation method performs the MC method, watershed transform, and then fuzzy clustering on size-weighted images blocks. It it claimed that this algorithm is robust in segmenting small objects, which plays an important role in remote-sensing image segmentation applications. Lu and Peng present in [5] a segmentation method based on improved FCM algorithm incorporating spatial information. Images to be segmented are first converted from RGB color image features into HIS space, and then an improved FCM clustering algorithm is used to get membership function of I component of pixels. An additional feature with the membership function values and H component of pixels is formed. Finally, a simple K-means method is applied for final clustering. A multistage method using hierarchical clustering for unsupervised image classification is presented [6]. In the first phase, the multistage method performs segmentation using a hierarchical clustering procedure which confines merging to spatially adjacent clusters and generates an image partition such that no union of any neighboring segments has homogeneous intensity values. In the second phase, the segments resulting from the first stage are classified into a small number of distinct states by a sequential merging operation. Cao et al. [7] propose a novel segmentation algorithm implemented within a level sets framework to deal with multi-class partitioning problem. Focusing on low resolution photographic gray aerial images, their method performs feature extraction at the first stage, evolves curves according to the proposed algorithm at the next stage and gets the partition result in the end. Wang et al. [8] propose a framework for incorporating spatial information with the aim of achieving robust and accurate segmentation in case of mixed noise without using experimentally set parameters. The proposed objective function has a new dissimilarity measure, and the weighting factor for neighborhood effect is fully adaptive to the image content.

3. One-class support vector machine

The basic idea of support vector machine (SVM) is to transform data points from data space to a high dimensional feature space using a kernel function so that the data points in the feature space become linearly separable. One-class SVM (OCSVM) [9] is a version of SVM for constructing a classifier using a set of positive training patterns. For a set of training patterns, OCSVM finds a hyperplane to separate these patterns from the origin O with maximum margin $\mathbf{w}\cdot\phi(\mathbf{x}_i)=\rho$ where $\mathbf{x}_i$ lies on the hyperplane H. The distance between the origin O and H is $\rho/\|\mathbf{w}\|$. To allow for the possibility of outliers in the data set and to make the method more robust, the projection values from a point on $\mathbf{w}$ need not be strictly larger than ρ, but the small projection value should be penalized. Therefore, slack variables ξ_i, $i=1, \ldots, l$, are introduced for small projection values, and the objective function and constraints are defined as

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2-\rho+\frac{1}{\nu l}\sum_{i=1}^{l}\xi_i, \\ s.t. \quad & \mathbf{w}\cdot\phi(\mathbf{x}_i)\ge\rho-\xi_i, \quad \xi_i\ge 0, \forall\mathbf{x}_i\in l, \end{aligned} \tag{1}$$

where ϕ is a nonlinear mapping from the input space to the feature space, and $\nu\in(0, 1]$ is a parameter which gives a trade-off between the maximum margin and the errors. When ν is small, the penalty on small projection values become substantial, thus few outliers should exist and the margin is small. On the contrary, when ν is large, many outliers with small projection values may exist and the margin is generally large. To solve the constrained optimization problem, Lagrangian is introduced. Also, to avoid working in the high-dimensional feature space, we pick a feature space where the dot product can be calculated directly using a kernel function K in the input space, and the Wolfe dual form of Eq. (1), which is a quadratic function in $\alpha_i' s$, becomes:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \\ s.t. \quad & 0\le\alpha_i\le\frac{1}{\nu l}, \quad \sum_{i=1}^{l}\alpha_i=1. \end{aligned} \tag{2}$$

This constrained optimization problem can be solved using a standard QP solver. Throughout this paper, the RBF kernel is adopted. That is:

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-q\|\mathbf{x}_i-\mathbf{x}_j\|^2) \tag{3}$$

where q is the width parameter. This equation implies that each input point is mapped on the octant surface of the unit ball in the high-dimensional feature space.

Many machine learning algorithms usually suffer from common disadvantages such as considerably long training time and overfitting. Most of the reasons of these disadvantages are from the fact that they employ the empirical risk minimization (ERM) principle to learn patterns. ERM minimizes risks resulting from the training data only; thus it may overfit a model. When a lot of training data, such as in an aerial image, are processed, the training time and performance become worse. Conversely, SVM uses the structural risk minimization (SRM) principle to train a model. SRM incorporates the model complexity into a learning process and may avoid the overfitting problem encountered in many learning algorithms. It is reported that SVM usually outperforms other soft-computing methods in efficiency, scalability, and performance [10,11]. For more details about SVM and OCSVM, please refer to [1,12].