Pattern Recognition 47 (2014) 3060–3069

Contents lists available at ScienceDirect

Pattern Recognition

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

Cooperative and penalized competitive learning with application to kernel-based clustering

Hong Jia [a], Yiu-ming Cheung [a,b,*], Jiming Liu [a]

[a] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[b] The United International College, BNU-HKBU, Zhuhai, China

ARTICLE INFO

Article history:
Received 19 September 2013
Received in revised form
5 February 2014
Accepted 13 March 2014
Available online 21 March 2014

Keywords:
Competitive learning
Cooperation mechanism
Penalization mechanism
Kernel-based clustering
Number of clusters

ABSTRACT

Competitive learning approaches with individual penalization or cooperation mechanisms have the attractive ability of automatic cluster number selection in unsupervised data clustering. In this paper, we further study these two mechanisms and propose a novel learning algorithm called Cooperative and Penalized Competitive Learning (CPCL), which implements the cooperation and penalization mechanisms simultaneously in a single competitive learning process. The integration of these two different kinds of competition mechanisms enables the CPCL to locate the cluster centers more quickly and be insensitive to the number of seed points and their initial positions. Additionally, to handle nonlinearly separable clusters, we further introduce the proposed competition mechanism into kernel clustering framework. Correspondingly, a new kernel-based competitive learning algorithm which can conduct nonlinear partition without knowing the true cluster number is presented. The promising experimental results on real data sets demonstrate the superiority of the proposed methods.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

As an efficient approach to clustering analysis, competitive learning has been widely applied to a variety of research areas such as data mining [1], computer vision [2], bioinformatics [3] and so forth. However, in an unsupervised learning environment, clustering problems can be extremely difficult especially when the number of clusters is unavailable in advance. That is because traditional methods, including the *k*-means algorithm [4] and Expectation-Maximization (EM) algorithm [5], need the users to specify the exact number of clusters as an input; otherwise, they will likely produce incorrect clustering results. Generally, choosing the cluster number is an ad hoc decision based on prior knowledge of given data and it becomes nontrivial when the data has many dimensions [6].

In the literature, competitive learning with different mechanisms have received wide attention due to their effectiveness and interpretability for automatic cluster number detection. For example, with a penalization mechanism, the Rival Penalized Competitive Learning (RPCL) [7] can automatically select the cluster number by gradually driving extra seed points (i.e. the variables that are learnable towards the center of data clusters in the input space) far away from the dense region of the input data set. In this learning approach, for each data observation (also called *input*), not only the winner among all seed points is updated to adapt to the input, but also the second winner is penalized by a much smaller fixed rate (also called *delearning rate* hereinafter). However, empirical studies have found that the performance of RPCL algorithm is sensitive to the delearning rate, whose optimal setting differs for variant problems [7,8]. To solve this problem, an improved version named Rival Penalization Controlled Competitive Learning (RPCCL) [8] was proposed, which controls the rival-penalized strength through an adaptive way based on the distance between the winner and the rival relative to the current input. In general, as pointed out in [9], both of RPCL and RPCCL always penalize the extra seed points even if they are much far away from the dense region of the input data set. Consequently, the learning curves of seed points obtained by these algorithms as a whole will not tend to convergence. By contrast, another variant of RPCL called Stochastic RPCL (S-RPCL) [9], developed from the Rival Penalized Expectation-Maximization (RPEM) algorithm [9], can lead to a convergent learning process by penalizing the nearest rival stochastically based on its posterior probability. Moreover, to theoretically analyze the convergence behavior of rival penalized leaning method, Ma and Wang [10] have presented a general form of RPCL algorithm, called distance-sensitive RPCL (DSRPCL) and proved that the correct convergence of DSRPCL is associated with

[*] Corresponding author at: Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China.
E-mail addresses: hongjia@comp.hkbu.edu.hk (H. Jia), ymc@comp.hkbu.edu.hk (Y.-m. Cheung), jiming@comp.hkbu.edu.hk (J. Liu).

http://dx.doi.org/10.1016/j.patcog.2014.03.010
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

the minimization of a cost function defined on the weight vectors of a competitive learning network. It has also been pointed out in [10] that the DSRPCL algorithm may result in wrong convergence when the specified cluster number k becomes much larger than the true cluster number k^* (i.e. $k>2k^*$). This phenomenon also exists in the other versions of the rival penalized learning algorithm as shown in [11]. In contrast, the Competitive Repetition Suppression (CoRe) clustering method proposed in [11] can give a good performance if it is initialized with sufficiently large number of seed points. This method is inspired by a cortical memory mechanism and extends the RPCL framework by allowing multiple winners and losers in each learning iteration.

Besides the penalization mechanism, a cooperation strategy can also be utilized for detecting the cluster number in the competitive learning paradigm. One example is the Competitive and Cooperative Learning (CCL) [12] algorithm, in which the winner of each learning iteration will dynamically cooperate with several nearest rivals to update towards the input data together. Consequently, the CCL can make all the seed points converge to the corresponding cluster centers and the number of those seed points stably locating at different positions is exactly the cluster number. Nevertheless, further empirical studies presented by Li et al. [13] indicate that the performance of CCL is somewhat sensitive to the initial positions of seed points. To overcome this difficulty, they have proposed an improved variant, namely Cooperation Controlled Competitive Learning (CCCL) method, in which the learning rate of each seed point within the same cooperative team is adjusted adaptively based on the distance between the cooperator and the current input. Nevertheless, some empirical studies have found that the CCCL algorithm may still not work well if the clusters are seriously overlapped or initial seed points are all gathered in one cluster.

To sum up, the competitive learning methods with a pure penalization or cooperation mechanism have different advantages and limitations. Therefore, designing a new competitive learning method, which features the merits of both penalization and cooperation mechanisms while counteracts their respective drawbacks, will definitely improve the accuracy of clustering analysis without knowing cluster number. However, since these two kinds of competitive mechanisms conduct an opposite learning process, i.e. penalization is to drive extra seed points far away from the input space while cooperation is to converge all seed points to the corresponding cluster centers, it is a nontrivial task to combine them into a single learning procedure. Moreover, all of the aforementioned competitive learning methods are based on the framework of k-means approach and only suitable for linearly separable clusters. Nevertheless, a nonlinearly separable cluster structure is common from a practical viewpoint. In the literature, kernel-based clustering methods have been widely used to analyze nonlinear clusters [14,15]. This kind of approaches utilizes kernel functions to map the original data into a high dimensional feature space, in which a linear partition will result in a nonlinear partition in the input space. Corresponding algorithms include the kernel k-means [16], global kernel k-means [17], kernel SOM [18,19] and so on. However, all these kernel clustering algorithms also need the number of clusters to be specified exactly, which becomes difficult in an unsupervised learning environment. To the best of our knowledge, conducting kernel-based clustering without knowing the cluster number has not been well studied yet.

In this paper, we further study the penalization and cooperation mechanisms, which actually conduct opposite shift tracks on the seed points (i.e. scatteration and aggregation), and explore a novel learning model which can simultaneously inherit the advantages of these two different kinds of mechanisms. Specifically, a new competitive learning algorithm, namely Cooperative and Penalized Competitive Learning (CPCL), which performs cooperation and penalization in a single competitive learning process is presented. In this method, given an input, the winner generated from the competition of all seed points will not only dynamically select several nearest competitors to form a cooperative team to adapt to the input together, but also penalize some other seed points which compete intensively with it. The cooperation mechanism here enables the closest seed points to update together and gradually converge to the corresponding cluster centers while the penalization mechanism supplies the other seed points with the opportunity to wander in the clustering space and search for more appropriate cluster centers. Consequently, this algorithm features the fast convergence speed and the robust performance against the initialization of seed points. Moreover, to handle the nonlinearly separable clusters, we further introduce the proposed cooperative and penalized competitive mechanism into the learning framework of the kernel k-means method. Correspondingly, a new kernel-based clustering algorithm which can nonlinearly partition given data without knowing the true cluster number is presented. Experiments on variant real data sets have demonstrated the good performance of proposed algorithms.

The rest of this paper is organized as follows. Section 2 describes the proposed CPCL approach and gives out the corresponding algorithm. Section 3 introduces the CPCL method into the kernel k-means framework and presents a new algorithm to solve the cluster number selection problem in kernel-based clustering. Then, Section 4 shows the experimental results on various real data sets. Finally, we draw a conclusion in Section 5.

2. Cooperative and penalized competitive learning (CPCL) approach

To simultaneously inherit the advantages of the two opposite competitive strategies (i.e., cooperation and penalization), we present a novel competitive learning model namely Cooperative and Penalized Competitive Learning (CPCL), which can perform cooperation and penalization in a single competitive learning process.

2.1. Cooperation and Penalization Mechanisms in CPCL

This sub-section describes the cooperation and penalization mechanisms of CPCL approach. Suppose we have N inputs, $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$, coming from k^* unknown clusters, and k $(k\geq k^*)$ seed points, $\mathbf{m}_1,\mathbf{m}_2,\ldots,\mathbf{m}_k$, which are randomly initialized in the input space. Subsequently, given an input $\mathbf{x}_t$ each time, the winner among k seed points denoted as $\mathbf{m}_c$ is determined by

$$I(j|\mathbf{x}_t)=\begin{cases}1 & \text{if } j=c=\arg\min\limits_{1\leq i\leq k}\{\gamma_i\|\mathbf{x}_t-\mathbf{m}_i\|^2\}\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

with the relative winning frequency γ_i of $\mathbf{m}_i$ defined as

$$\gamma_i=\frac{n_i}{\sum_{j=1}^{k}n_j}, \tag{2}$$

where n_i is the winning times of $\mathbf{m}_i$ in the past [20].

After selecting out the winner $\mathbf{m}_c$, the area centered at $\mathbf{m}_c$ with the radius $\|\mathbf{m}_c-\mathbf{x}_t\|$ is regarded as the territory of $\mathbf{m}_c$. Fig. 1 has shown the winner's territory in two-dimensional space. Any other seed points which have intruded into this territory will be dominated by $\mathbf{m}_c$. That is, any other seed point $\mathbf{m}_j$ which satisfies

$$\|\mathbf{m}_c-\mathbf{m}_j\|\leq\|\mathbf{m}_c-\mathbf{x}_t\| \tag{3}$$

will either cooperate with the winner or be penalized by it, such as $\mathbf{m}_1$, $\mathbf{m}_2$ and $\mathbf{m}_3$ in Fig. 1. Subsequently, a question is naturally arisen: how to design the cooperation and penalization mechanism for a seed point?

Download English Version:

https://daneshyari.com/en/article/530407

Download Persian Version:

https://daneshyari.com/article/530407

Daneshyari.com