Neurocomputing 273 (2018) 583–592

Contents lists available at ScienceDirect

Neurocomputing

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

A fast algorithm for nonsmooth penalized clustering

Ruizhi Zhou[a,b], Xin Shen[a,b], Lingfeng Niu[a,*]

[a] Research Center on Fictitious Economy and Data Science, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

A R T I C L E I N F O

Article history:
Received 13 July 2016
Revised 2 June 2017
Accepted 27 August 2017
Available online 7 September 2017

Communicated by Tao Li

Keywords:
BB method
Nonmonotone technique
Smoothing trust region
Nonsmooth penalized clustering

A B S T R A C T

As a novel framework of clustering analysis, penalized clustering is able to learn the number of clusters automatically, and therefore has aroused widespread interest recently. To address the computational difficulties arising from the nonsmoothness of the penalty, a simple iterative algorithm based on smoothing trust region (STR) can be used. However, since STR only needs first-order information of the model, it might exhibit slow convergence rate sometimes. To accelerate STR and further improve the efficiency of penalized clustering, we propose a nonmonotone smoothing trust region (NSTR) algorithm, in which nonmonotone technique and the Barzilai and Borwein (BB) method are utilized together. We also prove that the new algorithm is globally convergent and estimate its worst case computational complexity. Experimental results on both simulated and real-life data sets validate the effectiveness and efficiency of the proposed method.

1. Introduction

Cluster analysis is the process of partitioning a given set of data objects into different subsets based on some common properties of the data objects. Each subset is a cluster, such that points in a cluster are similar to one another, yet dissimilar to points in other clusters. It has been widely used in data mining, business intelligence, biological engineering, image processing, and social network [1–3]. A wide variety of clustering approaches have been proposed so far, including k-means, hierarchical clustering, spectral clustering and their variants [4–11]. Most of these clustering methods assume that the number of clusters is given beforehand. However, clustering is an unsupervised learning task, and estimating the number of clusters becomes a challenging problem in clustering analysis.

Recently, a novel framework for cluster analysis has been proposed [12–17], we call it as penalized clustering. In penalized clustering methods, the number of clusters can be learned from the training data automatically, rather than being required as a parameter. Let $\{x_i\}_{i=1}^n$ be a set of observations, μ_i represents the cluster centroid that covers x_i, where both x_i and μ_i are q-dimensional vectors. A general model of penalized clustering can be written in the following formulation:

$$\min_{\{\mu_i\}_{i=1}^n} \frac{1}{2}\sum_{i=1}^{n} L(x_i-\mu_i) + \lambda \sum_{i<j} R(\mu_i-\mu_j),$$

where the first term is a data fitting term, in which $L(\cdot)$ is a loss function that minimizes the distance between the observations and their corresponding centroids. The second term is a penalty term, in which $R(\cdot)$ is a regularization function that groups the close centroids together. $\lambda > 0$ is the model parameter that balances these two terms. In penalized clustering, the selection of regularization function $R(\cdot)$ directly affects the sparseness of centroids and the clustering results. In [12–14], the ℓ_p-norm ($p \geq 1$) based penalties are considered. These penalties are all formulated as convex. Since using a convex penalty may yield severely biased estimates, Pan et al. [15] suggested to utilize nonconvex penalties. Since then, penalized clustering methods with nonconvex penalties appeared in succession [16,17]. The penalties used by these methods include: the truncated lasso penalty(TLP) [18], the minimax concave penalty [19] and the $\ell_p (0 < p < 1)$ based penalty [17].

Solving the nonsmooth penalized clustering model may encounter some difficulties. One of the key difficulties is the presence of the nonsmooth penalty in its objective. Another difficulty is the inherent large-scale nature of the clustering problems. This renders the task of building fast and simple methods. In [17], we proposed an iterative algorithm based on smoothing trust region for ℓ_p-based penalized clustering and achieved good results. In this

* Corresponding author.
E-mail addresses: zhouruizhi13@mails.ucas.ac.cn (R. Zhou), shenxin14@mails.ucas.ac.cn (X. Shen), niulf@ucas.ac.cn (L. Niu).

http://dx.doi.org/10.1016/j.neucom.2017.08.048
0925-2312/

paper, we would like to extend this method to all the nonsmooth penalized clustering model. Noticing that the algorithm in [17] is designed to solve large-scale problems, only first-order information is used and the algorithm might exhibit slow convergence rate. Therefore, a natural question is how to devise schemes that remain simple, but can exhibit much better performance.

Many accelerated strategies for the first-order methods are presented in the literature. In 1983, Nesterov [20] presented an accelerated gradient method for solving a class of convex programmings. Nestrerov showed that, by this algorithm, the number of iterations to find a solution can be bounded by $O(1/\sqrt{\epsilon})$. This algorithm is shown to be the optimal among all methods having only first-order information at consecutive iterates [21]. Nesterov's accelerated gradient method has attracted much interest due to the increasing need to solve large-scale convex programming problems by using fast first-order method. Based on the presentation of Nesterov's method, fast gradient methods have been generalized and extended, see [21–28], for details. However, these methods are only applicable to convex situations. Because the regularization term in nonsmooth penalized clustering model may be nonconvex, Nesterov's acceleration strategies are not suitable for our problem.

The BB method, proposed by Barzilai and Borwein in [29], is another common used acceleration strategy. It often requires less computational work and speeds up the convergence greatly. The BB method together with the nonmonotone strategies does not require the problem to be convex, and can be applied to the nonconvex situation. Therefore, we plan to use the BB method to accelerate the STR algorithm. The BB method requires fewer storage space and very inexpensive computations. Consequently, the BB method has been generalized in many occasions. In [30], Raydan proposed an efficient global Barzilai and Borwein algorithm for large scale unconstrained optimization problems. The algorithm is combined with the traditional nonmonotone line search proposed by Grippo et al.[31] and further extended by Birgin et al. [32]. In [32], a nonmonotone projected gradient method is proposed for minimizing the differentiable functions on closed convex sets. [33] studied projected Barzilai–Borwein methods for large-scale box-constrained quadratic programming. In [34], Dai et al. developed a cyclic version of the Barzailai–Borwein gradient type method, which proves to have a better performance than the standard BB in many cases. More theoretical analysis, generalizations and variants of the BB method can be seen in [34–38] and the references therein.

The contributions of this paper are as follows.

- We propose a fast algorithm for nonsmooth penalized clustering problem. Smoothing trust region (STR) is utilized to handle the nonsmoothness of the regularization term. A nonmonotone smoothing trust region (NSTR) algorithm, in which the BB step and the nonmonotone technique are utilized together, is proposed to further accelerate the STR algorithm.
- We prove that the NSTR algorithm is globally convergent in theory and also estimate the worst case iteration complexity at the same time.
- Experiments are carried out on both synthetic and real-world data sets. Comparisons to several state-of-the-art clustering methods validate the effectiveness of both STR and NSTR. Comparison between STR and NSTR shows that NSTR is much faster than STR.

The paper is organized as follows. In the next section, some related works about smoothing approximation and trust region method are given. In Section 3, the new nonmonotone smoothing trust region method for nonsmooth penalized clustering is presented. In Section 4, we discuss the properties of this new algorithm, present its convergence and computational complexity analysis. Numerical studies are given in Section 5. Some conclusions are given in the last section.

2. Related work

2.1. Smoothing approximation

Smoothing approximation is one of the most effective ways to solve nonsmooth optimizations. The main idea of smoothing method is to use a sequence of parameterized smooth functions to approximate the original nonsmooth function. The smoothing function could be defined as follows:

Definition 1 [39]. Let f: $\Re^n \to \Re$ be a continuous function. $\tilde{f}: \Re^n \times \Re_+ \to \Re$ is called a smoothing function of f, if $\tilde{f}_\nu(\cdot)$ is continuously differentiable in $\Re^n$ for any fixed $\nu > 0$, and for any $x \in \Re^n$, $\lim_{z\to x, \nu\downarrow 0} \tilde{f}_\nu(z) = f(x)$.

The smoothing methods have been studied for decades [39–45]. Chen and Mangasarian [46] conduct a class of smooth approximations of the function $(t)_+$ by convolution. Let $\rho : R \to R_+$ be a piecewise continuous density function satisfying

$$\rho(s) = \rho(-s) \quad \kappa := \int_{-\infty}^{\infty} |s|\rho(s)ds < \infty.$$

Then

$$\phi(t, \nu) := \int_{-\infty}^{\infty} (t - \nu s)_+ \rho(s)ds,$$

from $R \times R_+ \to R_+$ is well defined. ϕ is a smoothing function of $(t)_+$ by Definition 1. Because many nonsmooth optimization problems can be reformulated by using the plus function $(t)_+$, the smooth approximations of these nonsmooth functions and their compositions can be defined by choosing a smooth approximation of (t_+). The framework of the smoothing method is presented in Algorithm 1.

Algorithm 1 The framework of smoothing method.

Initialization Given smoothing parameter ν_0, $\tilde{\epsilon}$ and corresponding constants, set $k = 0$.
while $\nu_k \geq \tilde{\epsilon}$ **do**
  Construct and solve the smoothing subproblem min $\tilde{f}_{\nu_k}(\mu)$,
  Update smoothing parameter ν_k according to some criterion,
  $k = k + 1$,
end while

2.2. Trust region method

In smoothing method, we need to solve a series of smooth subproblems with the given smoothing parameter ν. The trust region method is a popular choice to solve these smoothing subproblems [47]. It generates steps with the help of a quadratic model of the objective function in a region around the current iterate,

$$\Omega_k = \{\mu : \|\mu - \mu_k\| \leq \Delta_k\},$$

where Δ_k is the radius of Ω_k. The model is trusted to be adequate to the objective function in Ω_k, and a step to be the (approximate) minimizer of the quadratic model in trust region is chosen. The selected step is tested by some rules, if the step is acceptable, it is taken as the next iterate. If the step is not acceptable, it reduces the radius of trust region and finds a new minimizer. Different from the traditional line search method, the direction and length of the step are selected simultaneously. The framework of trust region is described in Algorithm 2. More details can be found in the monographs of [47,48].