IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2020, accepted March 12, 2020, date of publication March 17, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981429

A New DC Algorithm for Sparse Optimal Scoring Problem

GUO-QUAN LI[1], XU-XIANG DUAN[1], AND CHANG-ZHI WU[2]
[1]School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China
[2]School of Management, Guangzhou University, Guangzhou 510006, China

Corresponding author: Guo-Quan Li (ligq@cqnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11871128, in part by the Natural Science Foundation of Chongqing under Grant cstc2019jcyj-msxmX0282 and Grant cstc2019jcyj-msxmX0368, and in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJQN201900531.

ABSTRACT Linear discriminant analysis (LDA) has attracted many attentions as a classical tool for both classification and dimensionality reduction. Classical LDA performs quite well in simple and low dimensional setting while it is not suitable for small sample size data (SSS). Feature selection is an effective way to solve this problem. As a variant of LDA, sparse optimal scoring (SOS) with ℓ_0-norm regularization is considered in this paper. By using a new continuous nonconvex nonsmooth function to approximate ℓ_0-norm, we propose a novel difference of convex functions algorithm (DCA) for sparse optimal scoring. The most favorable property of the proposed DCA is its subproblem admits an analytical solution. The effectiveness of the proposed method is validated via theoretical analysis as well as some illustrative numerical experiments.

INDEX TERMS Linear discriminant analysis, sparse optimal scoring, ℓ_0-norm, dc algorithm.

I. INTRODUCTION

Linear discriminant analysis is a classical method for classification and dimensionality reduction in many applications, because of its simplicity, robustness, and predictive accuracy [1]. A test observation with predictor is classified to the class with centroid closest to the predictor, where distance is measured in the *Mahalanobis* metric using the pooled within-class covariance matrix. The observation is then assigned to the class having the maximum posterior class probability. LDA has gained considerable attention due to its generalization performance and it has been widely used in many real-world problems such as particle identification, epileptic detection [2], and decision making analytic [3], etc.

There are three different ways to tackle LDA, which are based on solving the normal model, Fisher's discriminant problem and the optimal scoring problem. For two-class problems, Mai and Zou [4] established the equivalence between the three methods which can be formulated as constrained versions of the Fisher's discriminant problem, the optimal scoring problem, and a least squares formulation of linear discriminant analysis, respectively, [4]–[6]. For simple and low-dimensional data, classical LDA enjoys quite well performance and it is known to fail for SSS datasets because the within-class covariance matrix of the features is singular. Consequently, sparse discriminant techniques have become popular due to their ability to provide increased interpretation as well as predictive performance for SSS data, since sparse classifier leads to easier model interpretation and may reduce overfitting of the training data.

Feature extraction and feature selection are two important techniques in dimensionality reduction of small samples with high dimensions in different application fields. Feature selection removes irrelevant features and redundancy to reduce the impact of the noise data and improve model interpretability by choosing the important features. Embedded method is one of feature selection method which integrates the search for an optimal subset of features into the classifier construction. A common strategy for embedded method is to use regularization term to induce sparsity with respect to input features, i.e., using norm constraints on the coefficient vector. This trick has been widely used for dimensionality reduction, such as graph based sparse feature extraction (OGSFE) [7], multi-class sparse discriminant analysis (MSDA) [8], sparse linear embedding (SLE) [9], discriminative low-rank preserving projection (DLRPP) [10], sparse principal component analysis (SPCA) [11], etc. In 2015, Mai and Zou [8] proposed MSDA based on the Bayes rule formulation of linear

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

discriminant analysis by using ℓ_1-norm penalty. Besides, all discriminant directions can be found at once by MSDA. Ames and Hong [12] proposed zero-variance sparse discriminant analysis approach which formulates the sparse discriminant analysis problem as an ℓ_1 penalty nonconvex optimization problem and discriminative directions can be found sequentially in the null-space of within-class scatter matrix. The sparse optimal scoring problem (SOS) was originally introduced by C. Clemmensen *et al.* in [5] for multi-class problem seeking at most $K-1$ sparse discriminant directions. There are two main differences between SOS and MSDA. The first one is that SOS imposes an elastic net penalty term (ℓ_1 plus ℓ_2-norm). The addition of ℓ_2-norm is beneficial to the prediction performance of SOS. Another difference is that SOS finds discriminant directions sequentially while MSDA seeks discriminant directions at once. A block coordinate descent method for SOS was proposed by using ℓ_1 regularization [5]. In [13], S. Atkins *et al.* suggested two kinds new numerical optimization schemes for solving the sparse optimal scoring formulation of LDA based on block coordinate descent, the proximal gradient method and the alternating direction method of multipliers. Meanwhile, the per-iteration costs of these methods were also discussed.

The most natural way to obtain sparse classifier is to use ℓ_0-norm in the regularization term. However, problem with ℓ_0-norm regularization term is NP-hard due to the discontinuous and combinatorial property of ℓ_0-norm. Hence, the general strategy used in above literatures is to replace ℓ_0-norm with ℓ_1-norm in order to circumvent this difficulty [5], [13], [14]. It's worth mentioning that Le Thi and Phan [15] studied the optimal scoring problem with ℓ_0-norm regularization term and SOS is solved alternatively by using two continuous nonconvex approximation of ℓ_0-norm. Motivated by [15], a new DC algorithm (DCA) will be proposed to solve sparse optimal scoring in this paper by using a suitable approximation of ℓ_0-norm. Different from the algorithm of [15], the subproblem of our new DCA is not only smooth but also admits an analytical solution.

The paper is organized as follows. Section II briefly dwells on sparse optimal scoring problem and proposes a new continuous nonconvex approximation of ℓ_0-norm and a block coordinate descent method by using the new approximation. In Section III, we state a new DC algorithm for the subproblem of the alternative schemes. In Section IV, we compare our new approximation of ℓ_0-norm with other related approximations. The main algorithm for SOS and its convergence properties are presented in Section V. The numerical experiments are reported in Section VI, and concluding remarks are given in Section VII.

II. SPARSE OPTIMAL SCORING PROBLEM

A. PROBLEM FORMULATION

Let X be an $n\times p$ data matrix, where the rows of X correspond to observations in R^p sampled from one of Q classes, ($Q\geq 2$). We assume that the data has been centered so that the sample mean is the zero vector. Optimal scoring generates a sequence of discriminant directions and conjugate scoring vectors as follows. Suppose that we have obtained the first $k-1$ discriminant vectors $w_1, w_2, \ldots, w_{k-1}\in R^p$ and scoring vectors $\theta_1, \theta_2, \ldots, \theta_{k-1}\in R^Q$. The kth discriminant vector w_k and scoring vector θ_k can be obtained by solving the following optimal scoring problem

$$\begin{aligned}&\min_{w_k,\theta_k}\ \|Y\theta_k - Xw_k\|_2^2\\ &s.t.\ \frac{1}{n}\theta_k^T Y^T Y\theta_k = 1,\\ &\quad \theta_k^T Y^T Y\theta_l = 0,\quad l=1,\ldots,k-1.\end{aligned}\tag{1}$$

where Y denotes the $n\times Q$ indicator matrix for class membership, defined by $Y_{ij}=1$ if the ith observation belongs to the jth class, and $Y_{ij}=0$ otherwise.

In this paper, we will study sparse optimal scoring problem which employs regularization via the combination of ℓ_0-norm and ℓ_2-norm, where the ℓ_0-norm is used for feature selection making model easy interpretate, and ℓ_2-norm may reduce overfitting of the training data. As before, suppose that we have identified the first $k-1$ discriminant vectors $w_1, w_2, \ldots, w_{k-1}\in R^p$ and scoring vectors $\theta_1, \theta_2, \ldots, \theta_{k-1}\in R^Q$. To calculate the kth sparse discriminant vector w_k and scoring vector θ_k, we solve the following sparse optimal scoring problem [15]

$$\begin{aligned}&\min_{w_k,\theta_k}\ \|Y\theta_k - Xw_k\|_2^2+\lambda_1\|w_k\|_2^2+\lambda_2\|w_k\|_0\\ &s.t.\ \frac{1}{n}\theta_k^T Y^T Y\theta_k = 1,\\ &\quad \theta_k^T Y^T Y\theta_l = 0,\quad l=1,\ldots,k-1.\end{aligned}\tag{2}$$

Here $\lambda_1\geq 0$ and $\lambda_2\geq 0$ are tuning parameters, and $\|w_k\|_0$ denotes the ℓ_0-norm of w_k, i.e. the number of nonzero elements of vector w_k. The optimization problem (2) is nonconvex because of nonconvex spherical constraints. Furthermore, problem (2) is NP-hard due to the presence of ℓ_0-norm.

B. NEW CONTINUOUS APPROXIMATION OF ℓ_0-NORM

In this section, we use a continuous nonconvex function to approximate ℓ_0-norm. For $\alpha>0$, let

$$\eta_\alpha(x)=\min\{1,\alpha x^2\},\quad \forall x\in R.\tag{3}$$

Then the nonconvex approximation of ℓ_0-norm is defined by

$$\|x\|_0\approx\sum_{i=1}^{n}\eta_\alpha(x_i).\tag{4}$$

Using the approximation (4), we can reformulate problem (2) in the form

$$\min_{(w_k,\theta_k)\in R^p\times\Omega^k}\ \|Y\theta_k - Xw_k\|_2^2+\lambda_1\|w_k\|_2^2+\lambda_2\sum_{i=1}^{p}\eta_\alpha(w_{ki})\tag{5}$$

where $\Omega^k = \{\theta_k \in R^Q : \theta_k^T D\theta_k = 1, \theta_k^T D\theta_l = 0, l = 1, \ldots, k-1\}$ and $D = \frac{1}{n}Y^T Y$.

It is not easy to solve formulation (5) efficiently due to the nonconvex set Ω^k although ℓ_0-norm has been approximated by a continuous function. The block coordinate descent method given by [5] will be adopted to iteratively approximate solution of problem (5). Specifically, suppose that we have an estimate (w^t, θ^t) of $(w_k\ \theta_k)$, at each iteration, we perform two steps alternately:

1. Fix $\theta_k = \theta^t$ and compute w^{t+1} by solving

$$w^{t+1} = \arg\min_{w\in R^p} \|Y\theta^t - Xw\|_2^2 + \lambda_1\|w\|_2^2 + \lambda_2 \sum_{i=1}^{p} \eta_\alpha(w_i), \tag{6}$$

2. Fix $w_k = w^{t+1}$ and compute θ^{t+1} by solving

$$\theta^{t+1} = \text{argmin}_{\theta\in\Omega^k} \|Y\theta - Xw^{t+1}\|_2^2. \tag{7}$$

An explicit solution of problem (7) can be found in polynomial time when w_k is fixed based on the following Lemma.

Lemma 1: [5] The problem (7) has optimal solution $\theta^{t+1} = s_k/\sqrt{s_k^T D s_k}$, where $s_k = (I - Q_{k-1}Q_{k-1}^T D)D^{-1}Y^T Xw^{t+1}$, and Q_{k-1} is the $Q \times (k-1)$ matrix whose columns are the previous $k-1$ solutions $\theta_1, \theta_2, \ldots, \theta_{k-1}$ consecutively.

Now, we are going to develop alternative schemes based DCA for SOS and present the convergence property of our method.

III. ALTERNATIVE SCHEMES FOR SOS

A. OUTLINE OF DC PROGRAMMING AND DCA

DC Programming and DCA address the problem of minimizing a function f which is a difference of convex functions on the R^n. Generally speaking, a so-called standard DC program takes the form

$$\alpha = inf\{f(x) := g(x) - h(x) | x \in R^n\} \qquad (P_{dc})$$

where g, h are lower semi-continuous proper convex functions on R^n. Such a function f is called a DC function, and $g-h$, a DC decomposition of f, while the convex functions g and h are DC components of f. Note that, the closed convex constraint $x \in C$ can be incorporated in the objective function of (P_{dc}) by using the indicator function on Cdenoted by X_C which is defined by $X_C(x) = 0$ if $x \in C$, and$+\infty$ otherwise.

For a convex function θ, the subdifferential of θ at $x_0 \in dom\theta$ is denoted by $\partial\theta(x_0) := \{y \in R^n | \theta(x) \geq \theta(x_0) + < x - x_0, y >, \forall x \in R^n\}$. The subdifferential $\partial\theta(x_0)$ generalizes the derivative in the sense that θ is differentiable at x_0 if and only if $\partial\theta(x_0) = \nabla\theta(x_0)$.

DCA is based on local optimality conditions and duality in DC programming The necessary local optimality condition for DC program (P_{dc}) at x^* is given by $\emptyset \neq \partial h(x^*) \subset \partial g(x^*)$. If $\partial g(x^*) \cap \partial h(x^*) \neq \emptyset$, then x^* is called a critical point of $g-h$, or a generalized Karush-Kuhn-Tucker point (KKT) of (P_{dc}). The main idea of DCA is simple: each iteration t of DCA approximates the concave part $-h$ by its affine majorization that corresponds to taking $y^t \in \partial h(x^t)$ and minimizes the resulting convex function. The generic DCA scheme can be described as follows:

Initialization: Let $x_0 \in R^n$ be an initial guess, $t = 0$.

Repeat

-Calculate $y^t \in \partial h(x^t)$

- Calculate $x^{t+1} \in \arg\min\{g(x) - < x, y^t > | x \in R^n\}$ (P_t)

Until Convergence.

DCA is a descent method (without linesearch) and has a linear convergence for DC programs. For more details on the convergence of DCA the reader is referred [16], [17]. DCA was first introduced especially for the standard DC program by Pham Dinh Tao in 1985. It has been successfully applied to a lot of different and various nonconvex optimization problems [17], [18].

B. DC FORMULATIONS AND DCA FOR (6)

To solve problem (5) by using the block coordinate descent method, we now focus on how to solve problem (6) efficiently. In this subsection, a new DC approach for problem (6) will be proposed by using suitable DC decomposition for the objective function of problem (6). In fact, the approximation $\eta_\alpha(x)$ can be rewritten as a DC function

$$\eta_\alpha(x) = g(x) - h(x), \tag{8}$$

where $g(x) = \alpha x^2, h(x) = -1 + \max\{\alpha x^2, 1\}$ are both convex functions defined on R. Then ℓ_0-norm of vector $x = (x_1, \ldots, x_p)^T \in R^p$ can be approximated by a DC function in the form

$$\|x\|_0 \approx \sum_{i=1}^{p} \eta_\alpha(x_i) = \alpha\|x\|_2^2 - \sum_{i=1}^{p} h(x_i). \tag{9}$$

Using the above DC decomposition, we can reformulate problem (6) as the following DC programming

$$\min_{w\in R^p} F_\theta(w) = G(w, \theta^t) - \lambda_2 H(w) \tag{10}$$

where $H(w) = \sum_{i=1}^{p} h(w_i)$ and $G(w, \theta^t) = \|Y\theta^t - Xw\|_2^2 + (\lambda_1 + \lambda_2\alpha)\|w\|_2^2$ are both convex functions, hence problem (10) is a standard DC program and can be iteratively solved by DCA. At each iteration, we need to calculate the subgradient $v^l \in \partial H(w^l)$ and solve the convex subproblem of the DCA scheme, namely

$$\min_{w\in R^p} \|Y\theta^t - Xw\|_2^2 + (\lambda_1 + \lambda_2\alpha)\|w\|_2^2 - \lambda_2\langle v^l, w\rangle. \tag{11}$$

Expanding the the objective of (11) and dropping the constant term show that (11) is equivalent to minimizing

$$w^T[X^T X + (\lambda_1 + \lambda_2\alpha)I]w - w^T(2X^T Y\theta^t + \lambda_2\, v^l).$$

Obviously, the objective function of subproblem (11) is strongly convex and it admits a unique solution over R^p. More precisely,

$$w^{l+1} = \frac{1}{2}[X^T X + (\lambda_1 + \lambda_2\alpha)I]^{-1}(2X^T Y\theta^t + \lambda_2 v^l). \tag{12}$$

By Sherman-Morrison-Woodbury Lemma, we have

$$[X^TX + (\lambda_1 + \lambda_2\alpha)I]^{-1} = M^{-1} - M^{-1}X^T(I + XM^{-1}X^T)^{-1}XM^{-1},$$

where $M = (\lambda_1 + \alpha\lambda_2)I$ and $I + XM^{-1}X^T \in R^{n\times n}$. Then an analytic solution of subproblem (11) could be given as follows

$$w^{l+1} = \frac{1}{2}[M^{-1} - M^{-1}X^T(I + XM^{-1}X^T)^{-1}XM^{-1}] \times (2X^TY\theta^t + \lambda_2 v^l). \quad (13)$$

The closed form (13) is better than (12) in terms of computational complexity, since it involves computing the inverse matrix of n order matrix $I + XM^{-1}X^T$ instead of p order matrix $X^TX + (\lambda_1 + \lambda_2\alpha)I$. Moreover, the inverse matrix of $I + XM^{-1}X^T$ is computed only once.

Now, we give a new DC algorithm for subproblem (6) which is summarized in Algorithm 1 as follows.

Algorithm 1 DCA for Subproblem (6)

Initialization: Let $l = 0$ and choose $w^0 \in R^p$.
Repeat

1. Compute $v^l \in \partial H(w^l)$.
2. Compute w^{l+1} by

$$w^{l+1} = \frac{1}{2}[M^{-1} - M^{-1}X^T(I + XM^{-1}X^T)^{-1}XM^{-1}] (2X^TY\theta^t + \lambda_2\, v^l).$$

3. $l = l + 1$.

Until Convergence.

In step 1, the subgradient of $H(w^l)$ can be calculated by

$$v_i^l = \begin{cases} 2\alpha w_i^l & \text{if } \alpha(w_i^l)^2 \geq 1 \\ 0 & \text{else .} \end{cases} \quad (14)$$

Since the global convergence of DCA is shown in [16] for a general problem setting including (6), the convergence property is also valid for Algorithm 1.

Theorem 1: Let $\{w^l\}$ be the sequence generated by Algorithm 1. The following statements hold.

(i) The sequence $\{F_\theta(w^l)\}$ is decreasing.

(ii) If the sequence $\{w^l\}$ is bounded, then every limit point of $\{w^l\}$ is a critical point of problem (6).

IV. SOME RELATED NONCONVEX APPROXIMATIONS FOR ℓ_0-NORM

Generally, there are three kinds of methods for treating ℓ_0-norm: convex approximation, nonconvex approximation and nonconvex exact reformulation. Till now, some nonconvex continuous approximation functions have been proposed for ℓ_0-norm, such as Capped-ℓ_1 approximation and piecewise exponential concave approximation. Recently, these two renowned nonconvex approximations have been successfully applied to SOS problem in [15], where the corresponding subproblem can be written as

$$\min_{w\in\mathbb{R}^p} \|Y\theta^t - Xw\|_2^2 + \lambda_1\|w\|_2^2 + \lambda_2\alpha\|w\|_1 - \langle v^l, w\rangle. \quad (15)$$

Obviously, problem (15) is nonsmooth and it is difficult and time consuming to solve this subproblem. On the contrary, the corresponding subproblem (11) of our new DCA (Algorithm 1) is not only smooth but also admits analytic solution.

In [15], the coordinate descent method is used to solve the nonsmooth subproblem (15), and the per-iteration computational costs of their method is $\mathcal{O}(\kappa np^2)$, where κ denotes the number of iterations of the coordinate descent method. As a comparison, we give a brief discussion on per-iteration computational costs of Algorithm 1. In step 1, $\mathcal{O}(p)$ floating point operations is required to compute $\partial H(w^l)$. In step 2, the inverse matrix of $I + XM^{-1}X^T$ can be computed at a cost of $\mathcal{O}(n^3)$ and w^{l+1} can be updated by using $\mathcal{O}(n^3 + n^2p)$ flops. Please note that $I + XM^{-1}X^T$ is independent of l and its inverse will be computed only once. Hence, when n is much smaller than p, then the per-iteration cost of our algorithm is $\mathcal{O}(p)$, i.e., the per-iteration cost of our approach scales linearly with the number of features of our data. Therefore, we can conclude that our algorithm is theoretically faster than the methods in [16].

V. MAIN ALGORITHM AND ITS CONVERGENCE PROPERTIES

In this section, we describe a new coordinate descent method based on DCA and investigate the convergence properties of our method which can be described as follows.

Algorithm 2 Alternating Scheme Based on DCA for Problem (5)

for $k = 1$ to K, compute $k-$th discriminant vector w_k as follows:
Initialization: Choose $w_k^0 \in R^p$. Let $\theta_k^0 = s_k^0/\sqrt{(s_k^0)^T D s_k^0}$, where $s_k^0 = (I - Q_{k-1}Q_{k-1}^T D)D^{-1}Y^TXw_k^0$.
Repeat

1. For fixed θ_k^l, compute w_k^{l+1} by Algorithm 1 using w_k^l as initialization.
2. For fixed w_k^{l+1}, compute

$$s_k^{l+1} = (I - Q_{k-1}Q_{k-1}^T D)D^{-1}Y^TXw_k^{l+1}$$

and set $\theta_k^{l+1} = s_k^{l+1}/\sqrt{(s_k^{l+1})^T D s_k^{l+1}}$.

3. $l = l + 1$.

Until convergence.
End for

Theorem 2: (i) Let $F(w_k, \theta_k)$ be the objective function of the problem (5). Then Algorithm 2 generates the sequence $\{(w_k^t, \theta_k^t)\}, k = 1, \ldots, K$, such that $\{F(w_k^t, \theta_k^t)\}$ is convergent.

(ii) If the sequence $\{(w_k^t, \theta_k^t)\}$ is bounded, then every limit point of this sequence is a critical point of the problem (5).

Proof: We first prove property (i). We assume that $\{(w_k^t, \theta_k^t)\}$ is generated by Algorithm 2. Note that w_k^{t+1} is a solution of (6) and w_k^l is also feasible for (6). Therefore, we get

$$F(w_k^t, \theta_k^t) \geq F(w_k^{t+1}, \theta_k^t).$$

On the other hand, θ_k^{t+1} is the solution of (7) with $w_k = w_k^{t+1}$, then we have

$$F(w_k^t, \theta_k^t) \geq F(w_k^{t+1}, \theta_k^t) \geq F(w_k^{t+1}, \theta_k^{t+1}).$$

Thus $\{F(w_k^t, \theta_k^t)\}$ is nonincreasing. Moreover $\{F(w_k^t, \theta_k^t)\}$ is convergent since $F(w_k, \theta_k)$ is nonnegative for all w_k and θ_k. This proves (i).

Now, we prove property (ii). Suppose that $\{(w_k^t, \theta_k^t)\}$ is bounded. Then there exists a convergent subsequence $\{(w_k^{t_j}, \theta_k^{t_j})\}$ and a pair (w_k^*, θ_k^*) such that $\{(w_k^{t_j}, \theta_k^{t_j})\} \to (w_k^*, \theta_k^*)$ as $j \to \infty$. We will prove that (w_k^*, θ_k^*) is a critical point of problem (5), i.e.,

$$\emptyset \neq \partial_{w_k} G(w_k^*, \theta_k^*) \cap \partial_{w_k} H(w_k^*), \tag{16}$$

$$\{\theta_k^*\} = \arg\min_{\theta_k \in \Omega_k} F_{w_k^*}(\theta_k). \tag{17}$$

Since $\{\theta_k^{t_j-1}\}$ is a subsequence of $\{\theta_k^t\}$, $\{\theta_k^{t_j-1}\}$ is also bounded. Without loss of generality, we can suppose that $\theta_k^{t_j-1} \to \theta_k^{**}$ as $j \to \infty$. Combining with

$$F(w_k^t, \theta_k^t) \leq F(w_k^t, \theta_k^{t-1}) \leq F(w_k^{t-1}, \theta_k^{t-1}),$$

we have

$$\lim_{t\to\infty} F(w_k^t, \theta_k^t) = \lim_{t\to\infty} F(w_k^t, \theta_k^{t-1}).$$

Using the fact that $F(w_k, \theta_k)$ is continuous, we have

$$F(w_k^*, \theta_k^*) = F(w_k^*, \theta_k^{**}).$$

Hence we can conclude that $\theta_k^* = \theta_k^{**}$, because problem (7) has a unique solution. Since $w_k^{t_j}$ is a solution of problem (6) with $\theta = \theta_k^{t_j-1}$, we deduce immediately from (ii) of Theorem 3.1 that

$$\emptyset \neq \partial_{w_k} G(w_k^{t_j}, \theta_k^{t_j-1}) \cap \partial_{w_k} H(w_k^{t_j}).$$

Therefore, there exists v^{t_j} such that

$$v^{t_j} \in \partial_{w_k} G(w_k^{t_j}, \theta_k^{t_j-1}) \cap \partial_{w_k} H(w_k^{t_j}).$$

Moreover, we have $v^{t_j} = \nabla_{w_k} G(w_k^{t_j}, \theta_k^{t_j-1}) = 2[X^T X + (\lambda_1 + \lambda_2)\alpha I]w_k^{t_j} - 2\,XY\theta_k^{t_j-1}$ since $G(w_k, \theta_k)$ is smooth.

Thus,

$$v^{t_j} \to v^* = 2[X^T X + (\lambda_1 + \lambda_2)\alpha I]w_k^* - 2\,XY\theta_k^*, \quad j \to \infty.$$

Consequently, invoking the definition of conjugate function and using Lemma 2 in [16], we have

$$v^* \in \partial_{w_k} G(w_k^*, \theta_k^*) \cap \partial_{w_k} H(w_k^*),$$

i.e., condition (16) holds.

From the step 2 in Algorithm 2, we have

$$\theta_k^{t+1} = s_k^{t+1} / \sqrt{(s_k^{t+1})^T D s_k^{t+1}}$$

where $s_k^{t+1} = (I - Q_{k-1}Q_{k-1}^T D)D^{-1}Y^T X w_k^{t+1}$. Let $t \to \infty$, we have

$$\theta_k^* = s_k^* / \sqrt{(s_k^*)^T D s_k^*}$$

here $s_k^* = (I - Q_{k-1}Q_{k-1}^T D)D^{-1}Y^T X w_k^*$. Thus condition (17) holds since problem (7) has a unique solution and this proves (ii).

VI. NUMERICAL EXPERIMENTS

In this section, to investigate the performance of the new proposed algorithm (called ℓ_0-DCA), we compare our algorithm with four state-of-the-art algorithms: ADCA proposed in [15] where the subproblem of the alternative schemes is also solved by DCA, and the three new methods for ℓ_1 regularized SOS based proximal method SDAP, SDAAP and SDAD which were proposed in [13]. Specifically, SDAP and SDAAP apply proximal gradient method and accelerated proximal method to solve the subproblem of the alternative schemes respectively, while SDAD employs the alternating direction method of multipliers to solve the subproblem of the alternative schemes. We use Python to implement our approach with the aid of Numpy a very popular open-source software library for numerical computation. All experiments are conducted on a personal computer with an Intel core i7-8750H CPU 8GB RAM. We train these algorithms to get suitable discriminant vectors, and predict the test observations by solving the problem:

$$\operatorname{argmin}_k \left\| \chi^T W - \mu_k^T W \right\|_2^2$$

where W is the linear transformation $W = [\omega_1, \omega_2, \omega_3, \ldots, \omega_k]$, ω_i $(i = 1, 2, 3, \ldots, k)$ are discriminant vectors.

As for the problem of selecting hyper-parameters, the standard 10-fold cross-validation technique is employed, where the parameters α, λ_1, and λ_2 are selected from the sets $\{1, 5, 10, 25, 50, 100, 200, 400\}$, $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ respectively. We take 10^{-10} as the value of the stop tolerance. The initial vector ω is ones vector. Furthermore, if $|\omega_{ki}| < 10^{-3}$ we will view ω_{ki} as an uncorrelated feature for $k = 1, 2, 3, \ldots K$.

A. DATASETS

The datasets which will be used for numerical experiment contain synthetic datasets and seven real world datasets collected from UCI or UCR Archive 2018. We train the models to get suitable discriminant vectors for the following data sets.

1) SYNTHETIC DATASETS

We first generate two Gauss simulation datasets to evaluate the effect to classification of our algorithm following the

strategy of [13]. Specifically, we obtain observations corresponding to the i_{th} class, $i = 1, 2, \ldots, K$, by sampling 30 observations from the Normal distribution with mean $\mu_i \in R^P$, if it's entries indexed by $100(i-1), \cdots 100i$, the elements value will equal to 0.7 and all remaining equal to 0, the covariance matrix $\sum \in R^{p\times p}$ constructed as follows.

• Data1: in this simulation process, all features are correlated with $\sum_{ij} = r$ for all $i \neq j$ and $\sum_{ij} = 1$ for all i. The experiment conducted by $K \in \{2, 4\}$, $r \in \{0, 0.1, 0.5, 0.7, 0.9\}$.

• Data2: in the second simulation process, $\sum$ is a block diagonal matrix with 100×100 blocks. For each pair of indices (i, j) in the same block we set $\sum_{ij} = r^{|i-j|}$, and set $\sum_{ij} = 0$ otherwise. As before, we repeat the experiment for each $K \in \{2, 4\}$, $r \in \{0, 0.1, 0.5, 0.7, 0.9\}$.

For each experiment, we sample 15 testing observations from each class. For each (K, r) pair we generate a dataset with 20 observations and use nearest centroid classification following projection onto the span of the discriminant directions to test ADCA, SDAP, SDAAP and SDAD. We use 10-fold cross validation to train the regularization parameters in the same environment.

2) REAL WORLD DATASETS

We also compare our classifier ℓ_0-DCA with the four renowned methods on seven benchmark datasets: Dbbodies, Dbsubjects, Beef, Coffee, ECG, Pen and OliveOil, in terms of classification accuracy, training time, and the number of selected features. These real world datasets are described as follows and the details are shown in Table 1.

TABLE 1. The details of the real world datasets.

No.	Dataset	Features	n_{train}	n_{test}	Classes
1	Dbbodies	4702	51	13	2
2	Dbsubjects	242	51	13	2
3	Beef	470	30	30	5
4	Coffee	286	28	28	2
5	ECG	136	23	861	2
6	Pen	3542	24	12	3
7	OliveOil	570	30	30	4

Dbbodies [19] and *Dbsubjects* [19] data sets consist of 64 e-mails from DBWorld news letter. Each attribute corresponds to a precise word or stem in the entire data set vocabulary. It is available from the link: http://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails.

Beef [21] data set consists of four classes of beef spectrograms, from pure beef and beef adulterated with varying degrees of offal. The data were first used in the time series classification literature by Bagnall *et al.* [20]. It is available from the link: http://www.timeseriesclassification.com/description.php? Dataset=Beef.

Coffee [22] data set is a two-class problem to distinguish between Robusta and Aribica coffee beans. The data were first used in the time series classification literature by Bagnall *et al.* [20]. It is available from the link: http://www.timeseriesclassification.com/description.php? Dataset=Coffee.

ECG data set is from a 67 year old male. The two classes correspond to two dates that the ECG was recorded. It is available from the link: http://www.timeseriesclassification.com/description.php?Dataset=ECGFiveDays.

Pen data set is multi-spectral images of three Penicillium species which was used in Clemmensen *et al.* [24]. It can be found in this acticle.

Each class of *OliveOil* [23] data set is an extra virgin olive oil from alternative countries. The data set can be found from the link: http://www.timeseriesclassification.com/description. php?Dataset=OliveOil.

B. EXPERIMENTS ON GUESS DATASETS

Table 2 shows the classification results of these five classifiers on synthetic data sets. In this table, the average percentage of accuracy of classifiers (ACC), the number of selected features (Feats), training time in seconds and their corresponding standard deviation over 10 trials are reported. From Table 2, we can see that (i) all classifiers get nice performance of classification. Moreover, our method get the best accuracy on both two data sets, which means that our classifier can select important features for classification. (ii) In terms of training time, our method is comparable. On Data1, the training time of ℓ_0-DCA is much less than that of other four methods.

C. EXPERIMENTS ON REAL WORLD DATASETS

We now compare our classifier with the four latest methods for sparse optimal scoring problem on seven benchmark data sets in terms of classification accuracy (ACC) and the number of selected features (Feats). In Table 3, the classification accuracy and the number of selected features for these five classifiers are listed, and the best results is shown by bold figure. From Table 3, we can get the following conclusions.

Accuracy: the accuracy of our method ℓ_0-DCA is 100.00, achieving the best results and is better than *ADCA* and *SDAD* on *Dbbodies* and *Coffee* datasets. SDAP and SDAAP work as well as ℓ_0-DCA on these two datasets. **Table 3** shows that ℓ_0-DCA attains better classification accuracy than other four methods on all data sets except *ECG*. On *ECG* dataset SDAP is slightly better than ℓ_0-DCA. SDAP and SDAAP perform best on 4/7 datasets. In a world, all of the five methods perform well on the seven real word datasets in term of accuracy.

Sparsity: as is shown in Table 3, the number of selected features in classification of our algorithm is less than some of the algorithms in *ADCA*, *SDAP*, *SDAAP* and *SDAD*. In term of the number of selected features, *ADCA* performs best on 6/7 dataset. The number of features selected by our algorithm is very close to that of ADCA which shows that ℓ_0-DCA also enjoy good feature selection capability.

Time: Figures 1-7 tell us that our method consumes less time than other four methods on all datasets we conduct. The time cost of the other four algorithms is several or even dozens of times that of our method on *Dbsubjects* and

TABLE 2. Comparing the classification performance with the latest methods ACDA, SDAP, SDAAP, SDAD on the synthetic datasets. Bold fonts indicates the best results in each row.

Dataset	Measures	ℓ_0-DCA	ADCA	SDAP	SDAAP	SDAD
Data1	Acc	**100.00 (0.64)**	96.32 (0.64)	**100.00 (0)**	**100.00 (0)**	98.93 (0.05)
	Feats	2.74 (0.48)	**2.01 (0.61)**	2.77 (0.35)	2.84 (0.74)	2.25 (0.14)
	Time	**0.12 (0.02)**	1.33 (0.12)	5.77 (2.11)	1.78 (0.22)	3.23 (0.04)
Data2	Acc	**97.78 (0.38)**	88.45 (0.57)	95.56 (0.37)	93.33 (0.12)	91.44 (0.81)
	Feats	25.50 (1.12)	**22.74 (1.87)**	28.25 (2.11)	27.75 (3.01)	32.75 (2.65)
	Time	25.35 (0.67)	33.27 (2.12)	89.47 (4.3)	**19.55 (1.12)**	28.84 (2.7)

TABLE 3. Comparing the classification performance with the latest methods ACDA, SDAP, SDAAP, SDAD on seven real world datasets. Bold fonts indicates the best results in each row.

Dataset	Measures	ℓ_0-DCA	ADCA	SDAP	SDAAP	SDAD
Dbbodies	Acc	**100.00(0)**	90.47(0.73)	**100.00(0)**	**100.00(0)**	91.43(0.41)
	Feats	35.00(1.01)	**15.02(0.22)**	27.33(0.83)	38.50(0.74)	34.33(16.24)
Dbsubjects	Acc	**85.71(0.28)**	**85.71(0.41)**	85.70(0.32)	**85.71(0.32)**	84.43(0.39)
	Feats	15.67(0.32)	**14.32(0.14)**	17.50(0.41)	28.67(0.22)	26.32(0.51)
Beef	Acc	**87.74(0.35)**	87.04(0.66)	80.00(0.51)	80.00(0.35)	78.85(0.88)
	Feats	35.56(0.64)	**28.49(0.73)**	48.11(1.42)	41.02(1.21)	38.73(0.89)
Coffee	Acc	**100.00(0)**	96.43(0.45)	**100.00(0)**	**100.00(0)**	96.42(0.21)
	Feats	16.24(0.42)	**14.61(0.35)**	17.12(0.15)	17.35(0.31)	20.22(0.47)
ECG	Acc	96.73(0.21)	97.09(0.76)	**97.55(0.28)**	97.31(0.43)	97.31(0.31)
	Feats	42.03(1.72)	35.12(1.06)	**9.33(0.34)**	21.67(2.01)	32.42(2.26)
Pen	Acc	**100.00(0)**	91.67(0.35)	91.67(0.34)	**100.00(0)**	83.33(0.72)
	Feats	98.50(6.74)	**90.63(2.23)**	223.45(2.17)	101.25(3.04)	51.50(3.32)
OliveOil	Acc	**96.67(0.71)**	95.23(0.33)	**96.67(0.26)**	96.67(0.32)	96.67(0.56)
	Feats	46.62(3.32)	**21.75(2.12)**	24.45(2.17)	52.24(2.78)	92.25(2.78)

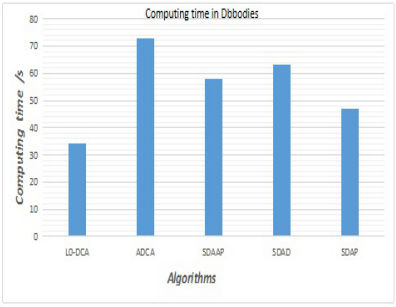

FIGURE 1. Training time on the Dbbodies dataset.

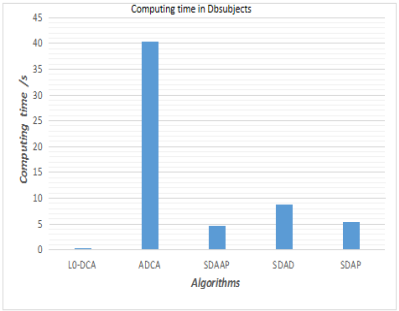

FIGURE 2. Training time on the Dbsubjects dataset.

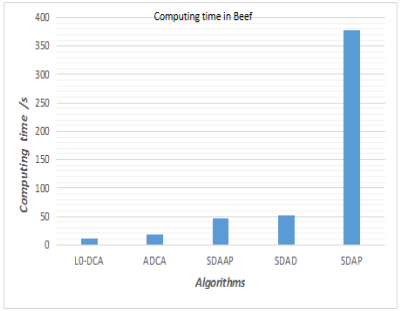

FIGURE 3. Training time on the Beef dataset.

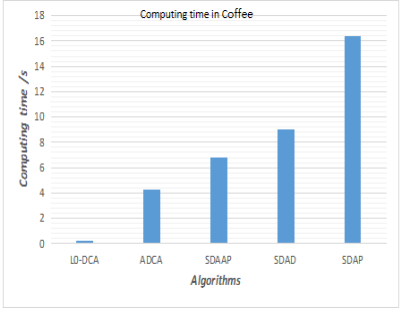

FIGURE 4. Training time on the Coffee dataset.

Coffee datasets. In general, our method ℓ_0-DCA outperforms the other four methods on these real datasets in term of training time. This can be explained by the fact that the computational resources required for each iteration scales linearly with the dimension of the data, since the subproblem in our method is smooth and admits closed form solutions.

Now, we want to illustrate the discriminant vectors can be used to visualize the datasets. The visualization of classification on *oliveoil* data set is shown in Figures 8-10, where the samples in each class are shown by using a distinct symbol. Figure 9 shows that the 1st discriminant vector and the 3rd

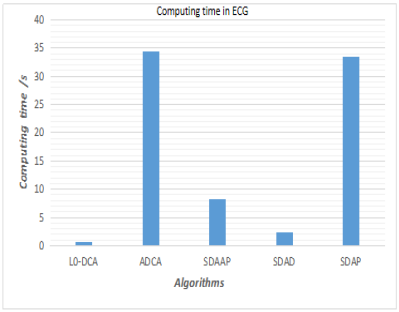

FIGURE 5. Training time on the ECG dataset.

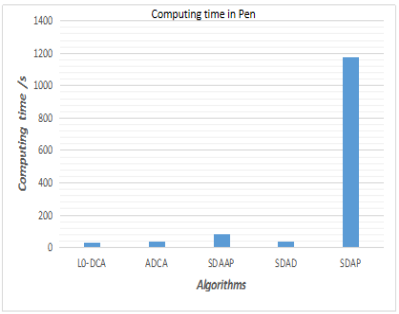

FIGURE 6. Training time on the Pen dataset.

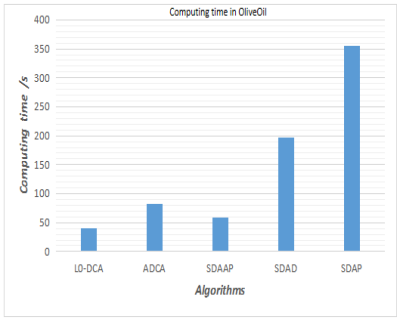

FIGURE 7. Training time on the OliveOil dataset.

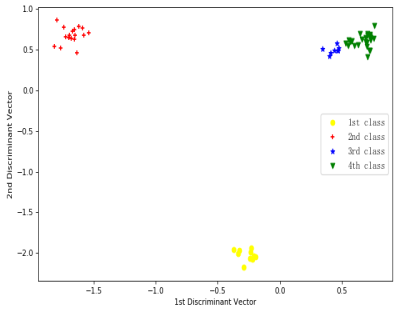

FIGURE 8. The oliveoil data is projected onto the 1st and the 2nd discriminant vectors.

discriminant vector obtained by our method could separate four classes samples very well.

Before the end of this section, we analyze the influence of three parameters to the accuracy of ℓ_0-DCA on all of the datasets used in this paper including synthetic datasets. We study the effect of different candidate value for one of the hyperparameters by fixing the other optimal parameters

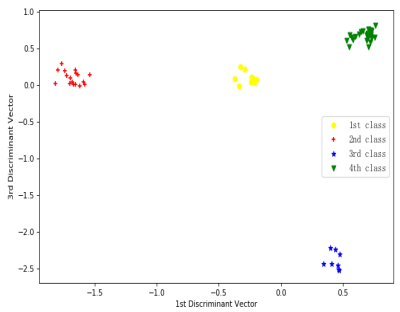

FIGURE 9. The oliveoil data is projected onto the 1st and the 3rd discriminant vectors.

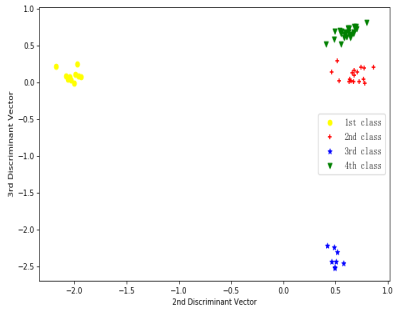

FIGURE 10. The oliveoil data is projected onto the 2nd and the 3rd discriminant vectors.

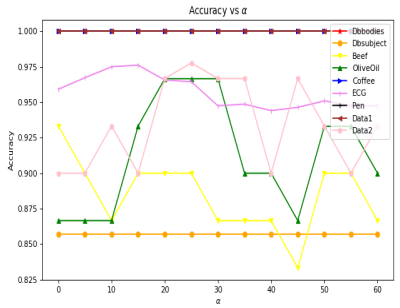

FIGURE 11. The influence of the different parameters α on the accuracy in ℓ_0-DCA.

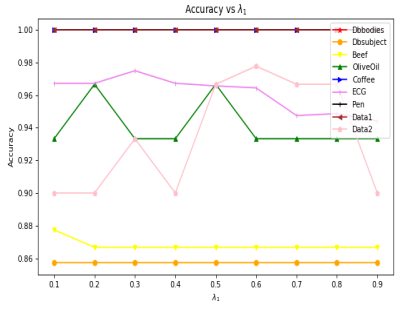

FIGURE 12. The influence of the different parameters λ_1 on the accuracy in ℓ_0-DCA.

to show its influence to accuracy. Figure 11 shows that accuracies attain the best results for most of datasets when $\alpha = 25$ although there are some fluctuation on several datasets. From Figures 11-12, we can find that on most datasets, accuracy is

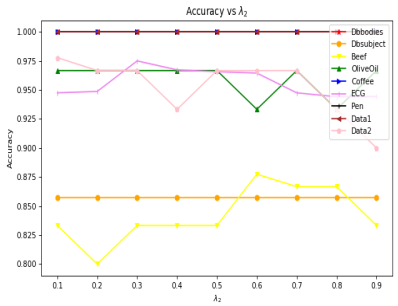

FIGURE 13. The influence of the different parameters λ_2 on the accuracy in ℓ_0-DCA.

not sensitive to the value of λ_1 and λ_2 except for the Oliveoil dataset, data2 and the ECG dataset.

VII. CONCLUSION

In this paper, we study sparse optimal scoring problem with ℓ_0 regularization. An alternating scheme based on DCA is proposed by using a nonconvex continuous function to approximate ℓ_0-norm and making suitable DC decomposition. One of the advantages of the new DCA is that its subproblems are smooth and have closed form solution at every iteration. The computational resources required for each iteration scale linearly with the dimension of the data. Moreover, we establish that any convergent subsequence of iterates generated by our algorithm converges to a critic point. Finally, experimental results show the efficiency of the proposed algorithm on accuracy and training time. For future work, we will focus on robust sparse optimal scoring problem with data uncertainty which affects classification accuracy and yields low quality solutions. How to find robust discriminant vectors for classification is our future work.

ACKNOWLEDGMENT

The authors would like to thank the referees for their valuable comments that have largely improved the presentation of this paper.

REFERENCES

[1] D. J. Hand,"Classifier technology and the illusion of progress," *Stat. Sci.*, vol. 21, no. 1, pp. 1–14, Feb. 2006.
[2] M. I. Khalid, T. Alotaiby, S. A. Aldosari, S. A. Alshebeili, M. H. Al-Hameed, F. S. Y. Almohammed, and T. S. Alotaibi, "Epileptic MEG spikes detection using common spatial patterns and linear discriminant analysis," *IEEE Access*, vol. 4, pp. 4629–4634, 2016, doi: 10.1109/access.2016.2602354.
[3] J. Kah Phooi Seng and K. Li-Minn Ang, "Big feature data analytics: Split and combine linear discriminant analysis (SC-LDA) for integration towards decision making analytics," *IEEE Access*, vol. 5, pp. 14056–14065, 2017, doi: 10.1109/access.2017.2726543.
[4] Q. Mai and H. Zou, "A note on the connection and equivalence of three sparse linear discriminant analysis methods," *Technometrics*, vol. 55, no. 2, pp. 243–246, May 2013.
[5] X.-D. Chen and H.-X. Lin, "Sparse discriminant analysis," *J. Comput. Appl.*, vol. 32, no. 4, pp. 1017–1021, Apr. 2013.
[6] M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, and X. Lin, "Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection," *Bioinformatics*, vol. 25, no. 9, pp. 1145–1151, 2008.
[7] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, "Structured optimal graph based sparse feature extraction for semi-supervised learning," *Signal Process.*, vol. 170, May 2020, Art. no. 107456, doi: 10.1016/j.sigpro.2020.107456.
[8] Q. Mai, Y. Yang, and H. Zou, "Multiclass sparse discriminant analysis," 2015, *arXiv:1504.05845*. [Online]. Available: http://arxiv.org/abs/1504.05845
[9] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 723–735, Apr. 2016.
[10] Z. Liu, J. Wang, G. Liu, and L. Zhang, "Discriminative low-rank preserving projection for dimensionality reduction," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105768.
[11] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Mar. 2010.
[12] B. P. W. Ames and M. Hong, "Alternating direction method of multipliers for penalized zero-variance discriminant analysis," *Comput. Optim. Appl.*, vol. 64, no. 3, pp. 725–754, Jul. 2016.
[13] S. Atkins, G. Einarsson, B. Ames, and L. Clemmensen, "Proximal methods for sparse optimal scoring and discriminant analysis," 2017, *arXiv:1705.07194*. [Online]. Available: http://arxiv.org/abs/1705.07194
[14] C. Leng, "Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data," *Comput. Biol. Chem.*, vol. 32, no. 6, pp. 417–425, Dec. 2008.
[15] H. A. Le Thi and D. N. Phan, "DC programming and DCA for sparse optimal scoring problem," *Neurocomputing*, vol. 186, pp. 170–181, Apr. 2016.
[16] P. D. Tao and L. T. H. An, "Convex analysis approach to D.C. Programming: Theory, algorithms and applications," *Acta Math. Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.
[17] H. A. Le Thi and T. Pham Dinh, "DC programming and DCA: Thirty years of developments," *Math. Program.*, vol. 169, no. 1, pp. 5–68, May 2018.
[18] C. Wu, C. Li, and Q. Long, "A DC programming approach for sensor network localization with uncertainties in anchor positions," *Journal of Industrial and Management Optimization*, vol. 10, no. 3, pp. 817–826, 2014.
[19] M. Filannino, "DBWorld e-mail classification using a very small corpus," in *Project of Machine Learning Course*. Manchester, U.K.: Univ. of Manchester, 201;.
[20] A. Bagnall, L. Davis, J. Hills, and J. Lines, "Transformation based ensembles for time series classification," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2012, pp. 307–318.
[21] O. Al-Jowder, E. K. Kemsley, and R. H. Wilson, "Detection of adulteration in cooked meat products by mid-infrared spectroscopy," *J. Agricult. Food Chem.*, vol. 50, no. 6, pp. 1325–1329, Mar. 2002.
[22] R. Briandet, E. K. Kemsley, and R. H. Wilson, "Discrimination of *Arabica* and *Robusta* in instant coffee by Fourier transform infrared spectroscopy and chemometrics," *J. Agricult. Food Chem.*, vol. 44, no. 1, pp. 170–174, Jan. 1996.
[23] H. S. Tapp, M. Defernez, and E. K. Kemsley, "FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils," *J. Agricult. Food Chem.*, vol. 51, no. 21, pp. 6110–6115, Oct. 2003.
[24] L. H. Clemmensen, M. E. Hansen, J. C. Frisvad, and B. K. Ersbøll, "A method for comparison of growth media in objective identification of Penicillium based on multi-spectral imaging," *J. Microbiolog. Methods*, vol. 69, no. 2, pp. 249–255, May 2007.

GUO-QUAN LI received the master's degree from the Department of Mathematics, Chongqing Normal University, China, in 2006, and the Ph.D. degree in operational research and cybernetics from Shanghai University, China, in 2009. He is currently an Associate Professor with the School of Mathematical Science, Chongqing Normal University. His research interests include data mining, machine learning, and optimization methods.

XU-XIANG DUAN is currently pursuing the degree with the Department of Mathematics Science, Chongqing Normal University, China. His research interests include data mining and machine learning.

CHANG-ZHI WU received the Ph.D. degree from Zhongshan University, China, in 2006. In 2006, he joined Chongqing Normal University as a Lecturer, where he was promoted as a Professor, in 2009. In 2013, he joined the Australasian Joint Research Centre for Building Information Modelling, Curtin University, as a Senior Research Fellow. He is currently a Professor with the School of Management, Guangzhou University. His main interests include both theoretical and practical aspects of optimization and optimal control and their applications in signal processing, civil engineering, and construction management.