

Optimality Conditions for Sparse Quadratic Optimization Problem

Duygu Üçüncü[1,2(✉)], Süreyya Akyüz[1], Erdal Gül[2], and Gerhard Wilhelm-Weber[3,4]

[1] Bahçeşehir University, 34353 İstanbul, Turkey
{duygu.ucuncu,sureyya.akyuz}@eng.bau.edu.tr
[2] Yildiz Technical University, İstanbul, Turkey
gul@yildiz.edu.tr
[3] Poznan University of Technology, Poznan, Poland
[4] Institute of Applied Mathematics, METU, 06800 Ankara, Turkey
gerhard.weber@put.poznan.pl

Abstract. Sparse models are preferred in machine learning problems because of their computational interpretability and it is seen in many applications such as in Google Page Rank, classification and regression problems, in the method of Principal Component Analysis (PCA) that finds the most important features and further applications in graphical models. In this study, we derive optimality conditions for the quadratic problem which has cardinality constraint imposing sparse solution. Our Quadratic model is a special application of ensemble pruning model in ensemble learning algorithms. Here, we refer to our previous study on this application in ensemble selection for clustering problems. The quadratic model proposed in this study optimizes trade-off between accuracy and diversity of discriminant functions (classifiers) simultaneously so that the best candidates of ensemble are selected for prediction step. The selection of the best classifiers in the ensemble is crucial for the overall performance of ensemble learning algorithms since redundant/outlier solutions in the ensemble library will decrease the overall prediction accuracy. In order to eliminate such candidates, both accuracy and diversity are taken into account when selecting the best subset of the ensemble. The cardinality constraint is further relaxed by considering various approximations such as l_1-norm regularization and student t-log likelihood approximations. Under these considerations and approximations, we build optimality criteria for our quadratic optimization problem with a cardinality constraint.

Keywords: Ensemble pruning · Machine learning
Sparse constraints · Difference of convex functions

© Springer Nature Switzerland AG 2019
H. C. Rodrigues et al. (Eds.): *EngOpt 2018 Proceedings of the 6th International Conference on Engineering Optimization*, pp. 766–777, 2019.
https://doi.org/10.1007/978-3-319-97773-7_67

1 Introduction

For many years, a simple approximate solution of an optimization problem is preferred rather than exact solution because they are easier to understand, and implement. Such solutions correspond to sparsity which means the solution contains a few nonzero components. Sparse optimization is an optimization that involve zero-norm which means cardinality, in objective or constraints.

Quadratic optimization problems arise naturally in several fields of applications. Sparse quadratic optimization is very preferred method in machine learning which is used in variety of applications such as signal processing, generalized eigenvalue problems (GEVs) [23], and ensemble selection [1].

In this paper we studied necessary and sufficient optimality conditions of general sparse quadratic and nonconvex approximations by using Difference of Convex (DC) programming theory. As an application we considered ensemble selection problem in machine learning and show that required optimality conditions are hold for this specified case. Furthermore as an application to the theory analyzed in this study, we implemented nonconvex approximation to zero-norm by using exponential term in [4] for ensemble selection problem. As far as we know, this approximation has not been used yet for ensemble cluster selection problem in the literature. From this aspect the optimization based analysis explained for ensemble selection problem together with an application is a novel contribution to ensemble selection literature in machine learning.

Traditional methods of ensemble clustering aim to use all clustering solutions to make a common decision. But the recent studies have shown that diverse and accurate clustering solutions within the ensemble enhance the general performance of ensemble learning. In order to select the best accurate and diverse subset among the ensemble library, ensemble selection methods are developed in the literature [2,6,8,9,13,30].

There are two important factors in selection of ensemble. These factors are the accuracy and diversity in solution of clustering ensemble. There exist a trade-off between accuracy and diversity of clustering solutions since we make concession of diversity while increasing accuracy, and vice versa. On the other hand, the optimum number of clusters in the clustering ensemble play an active role in the decision of the clustering ensemble. Recently, elimination of the unnecessary cluster solutions from the ensemble for finding of best ensemble clustering candidates has become one of the major problems in the literature.

In the next Sect. 2, background material for DC programming and its optimality conditions are given. In Sect. 3, we consider detailed information of an ensemble selection problem with its optimality conditions. Then, in Sect. 4, we present our implementations on different datasets from UCI machine learning repository [3]. Finally in the last section, we summarized the proposed approach in this study.

2 Mathematical Preliminaries

In this section, we explain some mathematical background exists in literature which will help our analysis.

Definition 1 [12]. *Let Γ be a convex set in $\mathbb{R}^n$. A real valued function $f : \Gamma \to \mathbb{R}$ is called a DC function on Γ, if there exist two convex functions $g, h : \Gamma \to \mathbb{R}$ such that f can be expressed in the form $f(x) = g(x) - h(x)$. DC programs are of the form $\min\{f_0(x) : x \in \Gamma, f_i(x) \leq 0\ (i = 1, \ldots, m)\}$, where $f_i = g_i - h_i\ (i = 1, \ldots, m)$.*

Let $\eta : \mathbb{R}^n \to \mathbb{R}$ be such that $\eta(x)$ is a lower semi-continuous and convex function and $\eta^*(y) = \sup\{x^T y - \eta(x) : x \in \mathbb{R}^n\}$ is the conjugate function of $\eta(x)$. Now we can see the relationship between DC programming and its dual problem by $\alpha = \inf\{g(x) - h(x) : x \in \mathbb{X}\} = \inf\{h^*(y) - g^*(y) : y \in \mathbb{Y}\}$, where $\mathbb{X}$ and $\mathbb{Y}$ are closed subsets of $\mathbb{R}^n$.

Taking into consideration α to be finite implies $dom\, g \subset dom\, h$ and $dom\, h^* \subset dom\, g^*$, where $dom\, g = \{x \in \mathbb{R}^n : g(x) < \infty\}$; i.e., f is finite whenever g is finite. A function is called a polyhedral convex function on $\mathbb{R}^n$ if its epigraph is the intersection of finitely many closed half spaces in $\mathbb{R}^{n+1}$ [20]. We can consider affine functions and indicator functions as examples for polyhedral convex functions. If $g(x)$ or/and $h(x)$ are polyhedral in DC programming then it provides required about optimality conditions and convergence of DC algorithm's convergence.

Let us define a sub-gradient of η at x^0 to be $\partial\eta(x^0) = \{y \in \mathbb{R}^n : \eta(x) - \eta(x^0) \geq (x - x^0)^T y, \forall x \in \mathbb{R}\}$. Herewith, $\partial\eta(\cdot)$ is a closed convex set. If $x^0 \in \text{int}(dom\,(\eta))$, then $\partial\eta(x^0) \neq \emptyset$.

In following, we have Generalized Karush-Kuhn-Tucker (KKT) condition:

(i) $\partial h(x^*) \subset \partial g(x^*)$ if x^* is an optimal solution to the (primal) DC programming.
(ii) $\partial g^*(y^*) \subset \partial h^*(y^*)$ if y^* is an optimal solution to the dual DC programming.

$\mathcal{P}$ and $\mathcal{D}$ being the solution sets of the primal and dual DC programming, we have

$$\bigcup\{\partial g^*(y) : y \in D\} \subset P \subset dom\, g, \bigcup\{\partial h(x) : x \in P\} \subset D \subset dom\, h^*. \tag{1}$$

For a sufficient local optimality condition we must state that x^* being a point that admits a neighborhood $U(x)$ such that

$$\partial h(x^*) \cap \partial g(x^*) \neq 0, \forall x \in U(x) \cap dom\, g \tag{2}$$

makes x^* a local minimizer of $g - h$.

Let us consider the following regularized nonconvex problem

$$\underset{x \in \mathcal{C}}{\text{minimize}}\ \ f(x) + \rho\|x\|_0. \tag{3}$$

Here, ρ is a positive regularization constant of penalty term of $\|x\|_0$, $\mathcal{C}$ is a convex set in $\mathbb{R}^n$ and f is a DC function on $\mathbb{R}^n$. Since $\|x\|_0$ refers to cardinality measure which can be defined as a step function; it can be approximated by a continuous function $z(x;\theta)$, where $\theta > 0$ is a controlling parameter of tightness of the approximation which leads us to a new form of problem (3)

$$\underset{x\in\mathcal{C}}{\text{minimize}} \; f(x) + \rho \sum_{i=1}^{n} z(x_i;\theta). \tag{4}$$

Assumption 1 [14]. *Let $z : \mathbb{R} \to \mathbb{R}$ be a family of functions of variable x and parameter θ satisfying the subsequent properties:*

(a) $\lim_{\theta\to\infty} z(x;\theta) = r(x)$, *for all* $x \in \mathbb{R}$.
(b) $z(x;\theta) = z(|x|;\theta)$ *for any* $\theta > 0$ *and for all* $x \in \mathbb{R}$, *and* $z(x;\theta)$ *is increasing on* $[0,\infty)$ *for the variable* x.
(c) For any $\theta > 0$, $z(x;\theta) = \phi(x;\theta) - \omega(x;\theta)$ $x \in \mathbb{R}$ *is a DC function, where both* $\phi(x;\theta)$ *and* $\omega(x;\theta)$ *are finite convex functions on* $\mathbb{R}$.
(d) $x\ell \geq 0$ *for all* $x \in \mathbb{R}$, $\ell \in \partial z(x;\theta)$ *where* $\partial z(x;\theta) = \{p-h : p \in \partial\phi(x;\theta),\, h \in \partial\omega(x;\theta)\}$.
(e) For any $a \leq b$ *and* $0 \notin [a,b]$, $\lim_{\theta\to\infty} \sup\{|y| : y \in \partial z(x;\theta),\, x \in [a,b]\} = 0$.

Once again, we get another equivalent form of problem (4), where $\Omega = \{(x,y) : (x,y) \in \mathcal{C}, |x_i| \leq y_i\, (i = 1,2,\ldots,n)\}$ *and using assumption (b), as follow:*

$$\underset{(x,y)\in\Omega}{\text{minimize}} \; f(x) + \rho \sum_{i=1}^{n} z(y_i;\theta). \tag{5}$$

Proposition 1 [14]. *A point $x^* \in \mathcal{C}$ is a global (resp. local) solution of the problem (4) if and only if $(x^*,|x^*|)$ is a global (resp. local) solution of problem (5). Moreover, if (x^*,y^*) is a global solution of (5) then, x^* is a global solution of problem (4).*

Theorem 1 [14]. *Let $\mathscr{M}$ and $\mathscr{M}_\theta$ be the solution sets of problems (4) and (5), respectively.*

(a) Assume that (θ_k) is a sequence of nonnegative numbers such that $\theta_k \to \infty$ and (x^k) is a sequence such that for any k, $x^k \in \mathscr{M}_{\theta_k}$. If $x^k \to x^*$, then $x^* \in \mathscr{M}$.
(b) If $\mathcal{C}$ is compact, then for any $\epsilon > 0$ there exist a $\theta(\epsilon) > 0$, such that for all $\theta \geq \theta(\epsilon)$ it holds: $\mathscr{M}_\theta \subset \mathscr{M} + B(0,\epsilon)$.
(c) If there exist a finite set $\mathcal{K}$ such that for all $\theta > 0$ $\mathscr{M}_\theta \cap \mathcal{K} \neq \emptyset$, then there exists $\theta_0 \geq 0$, such that for all $\theta \geq \theta_0$ we have: $\mathscr{M}_\theta \cap \mathcal{K} \subset \mathscr{M}$.

Corollary 1 [14]. *Suppose that z is concave on $[0,\infty)$, $\mathcal{C}$ is a polyhedral convex set having at least one vertex and f is concave, bounded below on $\mathcal{C}$. Then, Ω defined in problem (5), is also a polyhedral convex set, having at least one vertex. Let ϑ be the vertex set of Ω and $\overline{\mathscr{M}_\theta} = \{x : \exists y \in \mathbb{R}^n \text{ such that } (x,y) \in \vartheta\}$. Then, for all $\theta > 0$, it holds $\emptyset \neq \overline{\mathscr{M}_\theta}$ and there exists a $\theta_0 > 0$ such that for all $\theta \geq \theta_0$, we have $\overline{\mathscr{M}_\theta} < \mathscr{M}$.*

Observe that for the case of f being concave, bounded below on the polyhedral convex set $\mathcal{C}$ and z being an exponential approximation, the coherency between the original problem's solution and the approximated problem's solution has been studied by [4,19]. But [14] gives more strong results with Corollary 1, besides the general ones with Theorem 1. Our motivation here is to find the consistency between the original and the approximate solution; therefore, we need to define the characteristics of the local solutions of these problems.

Proposition 2 [14]. *Following properties are true.*

(a) $x^* \in \mathcal{C}$ is a local optimum of problem (3) if and only if x^* is a local optimum of the problem

$$\underset{x \in \mathcal{C}^*}{\text{minimize}}\ f(x), \tag{6}$$

where $\mathcal{C}^* = \{x \in \mathcal{C} : x_i = 0, \text{ for all } i \neq \text{supp}(x^*)\}$. Here $\text{supp}(x)$ refers to support of x.

(b) If $x^* \in \mathcal{C}$ is a local optimum of the problem (3), then for some $\overline{x^*} \in \partial f(x^*)$

$$\langle \overline{x^*}, x - x^* \rangle \geq 0 \quad \forall x \in \mathcal{C}^*. \tag{7}$$

Let us express problem (4) as

$$\underset{x}{\text{minimize}}\ G(x) - H(x) \tag{8}$$

Then, for a point $x^* \in \mathcal{C}$, condition (2) can be rewritten as

$$0 \in \partial G(x^*) - \partial H(x^*), \tag{9}$$

which is equivalent to

$$\langle \overline{x^*}, x - x^* \rangle + \langle \overline{y^*}, x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{C}, \tag{10}$$

for some $\overline{x^*} \in \partial f(x^*)$ and $\overline{y_i^*} \in \lambda \partial z(x_i^*; \theta)$, $\forall\, (i = 1, \ldots, n)$.

Thus, we are ready to give consistency results of local optimality.

Theorem 2 [14]. *Let $\mathfrak{T}$ and $\mathfrak{T}_\theta$ be the sets of $x \in \mathcal{C}$ satisfying the conditions (7) and (10), respectively. Then we can draw the following conclusions:*

(a) Assume that (θ_k) is a sequence of nonnegative numbers such that $\theta_k \to \infty$ and (x^k) is a sequence such that for any k, $x^k \in \mathfrak{T}_\theta$. If $x^k \to x^*$, then $x^* \in \mathfrak{T}$.

(b) If $\mathcal{C}$ is compact, then for any $\epsilon > 0$ there exists a number $\theta(\epsilon) > 0$, such that for all $\theta \geq \theta(\epsilon)$, it holds $\mathfrak{T}_\theta \subset \mathfrak{T} + B(0,\epsilon)$.

(c) If there exists a finite set $\mathcal{K}$ such that for all $\theta > 0$, $\mathfrak{T}_\theta \cap \mathfrak{T} \neq \emptyset$ is fulfilled, then there exists number $\theta_0 \geq 0$ such that for all $\theta \geq \theta_0$, it holds $\mathfrak{T}_\theta \cap \mathcal{K} \subset \mathfrak{T}$.

3 An Application: DC Approximation to Sparse Ensemble Selection Problem

Sparsity makes it possible to represent and reconstruct high-dimensional data with a smaller dimension, to make computation faster and to enable extremely large-scale calculations. Penalizing the objective function with the l_0-norm, which means cardinality, is the most used way of imposing sparsity in optimization problems. Sparse optimization problems have been employed in machine learning for years. Studies involving the zero-norm can be solved by convex optimization, nonconvex optimization and nonconvex exact reformulation with proper approximations. For convex approximation, one of the well-known approach is called LASSO (Least Absolute Shrinkage and Selection Operator), which replace the l_0-norm by the l_1-norm [27]. The second category involves in approximating $\|x\|_0$ by a nonconvex continuous function, e.g., an exponential concave function [4], or a logarithmic function [28]. The last category which is a nonconvex exact reformulation by a continuous nonconvex program is mostly used in machine learning applications such as feature selection in SVM [15]. Because of its intractability for large-scale datasets, an exact penalty method is proposed by DC programming in [18,26].

DC approximation problems have many applications in machine learning, including feature selection, ensemble selection, generalized eigenvalue problem, l_1-SVM, l_∞-SVM, etc. For instance, in [23] generalized eigenvalue (GEV) problem is formulated as a sparse DC programming and solved by majorization-minimization (MM) approach. In their study, the authors considered a special application of sparse GEVs problem on canonical correlation analysis (CCA) which is implemented on cross-language document retrieval. In [14], sparse DC Algorithms (DCA) method is used to solve l_1-SVM which is known as one of the most preferred methods.

In this study, we consider ensemble selection problem in machine learning specifically for clustering problems. As it is stated in Sect. 1, the prediction performance of ensemble learning is better than single clustering algorithm.

In the following subsection, we describe ensemble cluster selection problem as an application to sparse quadratic optimization problem.

3.1 Ensemble Cluster Selection

Clustering which is one of the popular unsupervised learning method in machine learning is based on grouping similar objects according on their proximity in vector space. The goal of clustering is to group objects which have joint features inside of the group while having dissimilarities with other groups. There exist lots of studies about clustering algorithms in the literature [7,10,16]. One of the popular clustering algorithm is k-means [10] which is centroid-based clustering.

It is empirically shown in the literature that community of clustering solutions give better performance than a single clustering solution which is called in general ensemble learning. An ensemble clustering is a set of clustering solutions

created by different clustering methods, different sampling of the data or different parameters of a clustering method [2,8,11]. There exists a number of studies showing that the performance of a learning algorithm can be increased by using ensemble learning (e.g., bagging [5], boosting [21], stacking [29]).

In this section, we will consider a model (11) which optimizes the accuracy and diversity trade-off with the help of the idea given in [31]

$$\begin{aligned} &\text{maximize } x^T \widetilde{D} x \\ &\text{subject to } \textstyle\sum_{i=1}^{n} x_i = k, \\ &x_i \in \{0,1\} \ (i = 1, 2, \ldots, n) \end{aligned} \tag{11}$$

where $\widetilde{D}$ is the normalization of matrix D, determined by the same idea as in [31]. Subsequently, we give the definition of Normalized Mutual Information (NMI) function [24], which measures the similarities between each clustering solution where, used to determine each entry of matrix D:

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} i, \tag{12}$$

$$NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}. \tag{13}$$

In Eqs. (12) and (13), $I(X,Y)$ is the mutual information function between clusters X and Y, $p(x)$ and $p(y)$ are marginal distribution functions and $H(X)$, $H(Y)$ are entropy functions corresponding to clusters X and Y. Let $E = \{C_1, C_2, \ldots, C_n\}$ be the library of different clustering solutions, $NMI(C, C_i)$ be the function given in Eq. (13) and $SNMI(C, E)$ be a function that measures the accuracy given as follows:

$$SNMI(C,E) = \sum_{i=1}^{n} NMI(C, C_i). \tag{14}$$

In this case, matrix D is defined in this way:

$$D_{ij} = \begin{cases} D_{ii} = SNMI(C, C_i) \quad (i = 1, \ldots, n), \\ D_{ij} = 1 - NMI(C_i, C_j) \quad (j = 1, \ldots, n). \end{cases} \tag{15}$$

Here, the non-diagonal elements of the matrix D measure the diversity between two solutions while the diagonal elements measure the quality of the i-th solution. Let us consider the problem (11). It is clear that problem (11) is a special case of sparse GEV problem where $x^T B x = 1$, B is positive definite matrix, is missing. In order to make ensemble selection problem sparse, the constraint ($x^T x = 1$) is added to problem (11), where B is taken as identity matrix.

It is very important to relax the problem (11) by using Lagrangian Relaxation where ρ is penalization parameter. Hence, the constraint $\sum_i^n x_i = k$ can be written as $\|x\|_0 = k$ which corresponds to the cardinality of the subset selected

from the ensemble. That is, we use sparsity to relax the binary vector x to the real vector. Thus, problem (11) turns out as the following constrained optimization program:

$$\begin{aligned} &\underset{x\in R^n}{\text{maximize}}\ x^T\widetilde{D}x-\rho\|x\|_0 \\ &\text{with respect to } x^Tx=1. \end{aligned} \tag{16}$$

Let us consider problem (16) where $\widetilde{D}$ is from a class of symmetric and nonnegative matrices. Then problem (16) means a maximization of nonconcave objective function over a convex constraint set. $A := \{x : x^Tx = 1\}$ is NP-hard. Since the constraint $x^Tx = 1$ restricts the solution space more than it is required, we relaxed our equality constraint to $x^Tx \leq 1$. Thus problem (16) turns into

$$\begin{aligned} &\underset{x\in R^n}{\text{maximize}}\ x^T\widetilde{D}x-\rho\|x\|_0 \\ &\text{with respect to } x^Tx\leq 1. \end{aligned} \tag{17}$$

Let us take the problem (17) to its standard form as follows:

$$\begin{aligned} &\underset{x\in R^n}{\text{minimize}}\ -x^T\widetilde{D}x+\rho\|x\|_0 \\ &\text{with respect to } x^Tx\leq 1. \end{aligned} \tag{18}$$

Let $\mathbb{S}^n$ be a set of positive semidefinite matrices and $\tau \in \mathbb{R}$. Then we can say that $\widetilde{D}+\tau I_n \in \mathbb{S}^n$, since if $\tau \geq 0$ then $\widetilde{D} \in \mathbb{S}^n$. If $\widetilde{D}$ is indefinite we can choose $\tau \geq -\lambda_{min}(\widetilde{D})$ to make certain that again $\widetilde{D}+\tau I_n \in \mathbb{S}^n$. Thus, we can say $\widetilde{D}+\tau I_n \in \mathbb{S}^n$ is true for any $\widetilde{D} \in \mathbb{S}^n$ if $\tau \geq \max(0, -\lambda_{min}(\widetilde{D}))$ [17]. Therefore, problem (18) can be written as

$$\begin{aligned} &\underset{x}{\text{minimize}}\ \tau\|x\|_2^2-\left[x^T\left(\widetilde{D}+\tau I\right)x-\rho\|x\|_0\right] \\ &x^Tx\leq 1. \end{aligned} \tag{19}$$

In this paper we choose one of the well-known nonconvex approximation to zero-norm introduced in [4] by

$$z(x;\theta)=1-e^{-\theta|x|}. \tag{20}$$

Observe that $z(x;\theta)$ is a concave function that can be written as a DC function given by

$$z(x;\theta)=\theta|x|-\left(\theta|x|-(1-e^{-\theta|x|})\right), \tag{21}$$

which satisfies Assumption 1, part (c).

Thus, problem (19) can be formulated as

$$\begin{aligned} &\underset{x}{\text{minimize}}\ \tau\|x\|_2^2-\left[x^T\left(\widetilde{D}+\tau I\right)x-\rho\sum_{i=1}^n\left[\theta|x_i|-\left(\theta|x_i|-(1-e^{-\theta|x_i|})\right)\right]\right] \\ &\text{with respect to } x^Tx\leq 1. \end{aligned} \tag{22}$$

Note that objective function in (22) involves absolute value functions which is not differentiable; thus, problem (22) can be reformulated by using Assumption 1, part (b). We introduce a new variable y such that problem (22) turns into

$$\underset{x\in\Omega}{\text{minimize}}\, \tau \|x\|_2^2 - \left[x^T \left(\widetilde{D} + \tau I\right) x - \rho \sum_{i=1}^n \left[\theta y_i - \left(\theta y_i - (1 - e^{-\theta\, y_i})\right)\right]\right] \quad \text{with respect to } x^T x \leq 1, \tag{23}$$

where $\Omega = \{(x, y) : (x, y) \in \mathcal{C}, |x_i| \leq y_i\, (i = 1, 2, \ldots, n)\}$. Observe that problem (22) and problem (23) are equivalent by Proposition 1.

Next we will show that problem (23) satisfies optimality conditions and necessary assumption given in Sect. 2. Hence, this new formulation corresponds to an application of Eq. (5) with $f(x) = \tau \|x\|_2^2 - x^T \left(\widetilde{D} + \tau I\right) x$ and $z(y;\theta) = \left[\theta y_i - \left(\theta y_i - (1 - e^{-\theta\, y_i})\right)\right]$. Here, $\theta y_i := \phi(y_i;\theta)$ and $\theta y_i - (1 - e^{-\theta y_i}) := \omega(y_i;\theta)$ given in Assumption 1, part (c), where $\phi(y;\theta)$ and $\omega(y;\theta)$ are finite convex functions on $\mathbb{R}$. Thus our $z(y;\theta)$ is DC function. Let $\mathscr{M}$ and $\mathscr{M}_\theta$ be the solution sets of problem (4) and problem (5), respectively. It is proved (on Theorem 1) that if (θ_k) is a sequence of nonnegative numbers such that $\theta_k \to \infty$, (x^k) is a sequence such that $x^k \in \mathscr{M}_{\theta_k}$ for any k and $x^k \to x^*$, then $x^* \in \mathscr{M}$. Since feasible region given by $x^T x \leq 1$ together with the inequality constraint $-y_i \leq x_i \leq y_i$ is compact, for any $\epsilon > 0$ there is a number $\theta(\epsilon) > 0$ such that $\mathscr{M}_\theta \subset \mathscr{M} + B(0, \epsilon)$, $\forall \theta \geq \theta(\epsilon)$ by using Theorem 1, part (b). It is important to note that result of Theorem 1 guarantees the convergence to a solution of problem (4) for any choice of a nonnegative parameter θ of a nonconvex approximation $z(x;\theta)$. Moreover, by part (b) of Theorem 1, any optimal solution of problem (5) is in an ϵ-neighborhood of optimal solution for problem (4).

It is clear that our nonconvex approximation z is concave on $[0, \infty)$, f is concave bounded below on $\mathcal{C}$, and the feasible region of problem (23) is polyhedral in some piecewise manner since the constraint $-y_i \leq x_i \leq y_i$ contribute partially to a polyhedral nature of the convex region. However, since the disk condition $x^T x \leq 1$ are non-polyhedral we can not in general say that the feasible set is polyhedral but we can say the feasible set is polyhedral along it's boundary. Furthermore, the constraint $x^T x \leq 1$ is elementary constraint since it is a section in a second rather than first order cone and usually it does not lead to any further complications. Then by Corollary 1, there exists the set $\overline{\mathscr{M}_\theta} = \{x : \exists y \in \mathbb{R}^n \text{ such that } (x, y) \in \vartheta\}$, where ϑ is a vertex set of Ω. Then, $\emptyset \neq \overline{\mathscr{M}_\theta}$, $\forall \theta > 0$ and there exist $\theta_0 > 0$ such that for all $\theta \geq \theta_0$ it holds $\overline{\mathscr{M}_\theta} < \mathscr{M}$.

Let us consider (23). In order to represent the terms in problem (23) as in Proposition 2 of Sect. 2 we have the following definitions.

$G(x) := \tau \|x\|_2^2 + \rho \sum_{i=1}^n \theta y_i$,

$H(x) := \left[x^T(\widetilde{D} + \tau I)x + \rho \sum_{i=1}^n (\theta y_i - (1 - e^{-\theta y_i}))\right]$,

which leads us to Theorem 2. Whenever necessary optimality conditions given by Eqs. (7) and (10) are satisfy, then Theorem 2 is hold.

4 Numerical Experiments

Convex-Concave Programming (CCP) can also be viewed as a version of DC algorithms [25]. As a start-up implementation the ensemble selection problem (19) for clustering [1] is implemented by using CVXPY library [22] with non-convex approximation $\|x\|_0 := 1 - e^{-\theta|x|}$. We used different datasets from UCI Machine Learning Repository [3], where we explained the properties of those datasets by Table 1. The first column of Table 1 represents data names while the second column represents characteristics of the dataset which are called Multivariate and Time-Series. The third column gives information about characteristics of attributes. These are Integer, Real and Categorical. Fourth, fifth and sixth columns show the number of instances, the number of attributes and the number of clusters information, respectively.

Table 1. Data set information.

Data	DataChar	AttChar	NumIns	NumAtt	NumCls
WINE	Multivariate	Integer, real	178	13	3
SYNT	Time-series	Real	600	N/A	6
SEG	Multivariate	Real	2310	19	7
ECOLI	Multivariate	Real	336	8	8
GLASS	Multivariate	Real	214	10	6
ZOO	Multivariate	Categorical, Integer	101	17	7

Table 2. (NMI) values of DCCP algorithm.

Data	DCCP	Joint criterion	Cluster select	Cardinality of sparse set
Wine	**0.781696**	0.716708	0.566190	52
Synthetic control	**0.832222**	0.698854	0.761608	55
Segmentation test	**0.461336**	0.396007	0,197260	55
Ecoli	0.578387	**0.634155**	0.494483	55
Glass	**0.714068**	0.668443	0.660236	54
Zoo	**0.735905**	0.713078	0.665306	55

The performance results are evaluated by using (NMI) values, since the problem we consider is unsupervised so that the value of mutual information within the clusters reflects the quality of clustering decision. In other words, the closer the (NMI) values are to 1, the better the quality of clustering results. We compared our results with greedy based approaches in [8], which are called Joint Criterion and Cluster Select. Joint Criterion Method unite quality and diversity terms into a joint criterion function while Cluster Select Method takes in consideration every clustering solution in the ensemble as an entity and it analyzes the relationship of these clustering solutions [8]. The last column of

Table 2 shows the cardinality of ensemble selection. In other words, they refer to cardinality of sparse solution. It is clear that almost for each dataset convex concave programming gives better performance in terms of (NMI) values than other greedy based methods.

5 Conclusion

In this study, we compared optimization versus greedy based ensemble selection algorithms by using DC programming. As it stated in Sect. 4, performance of optimization based solution by DC algorithm is better than greedy methods in [8]. Furthermore, we analyzed optimality conditions of ensemble selection model by using supportive theorems and conclusions given in [14]. We also show that ensemble selection model satisfies necessary assumptions required for convergence to optimal solution.

Notes and Comments. This study is a part of PhD Dissertation of Doctoral Programme of Yildiz Technical University and also it is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) Project no:116E064.

References

1. Akyüz, S., Otar, B.: Ensemble clustering selection by optimization of accuracy-diversity trade off. In: 2017 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2017)
2. Azimi, J., Fern, X.: Adaptive cluster ensemble selection. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 992–997 (2009)
3. Bache, K., Lichman, M.: UCI Machine Learning Repository (2013)
4. Bradley, P.S., Mangasarian, O.L., Rosen, J.B.: Parsimonious least norm approximation. Comput. Optim. Appl. (1998)
5. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
6. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: Twenty-First International Conference on Machine Learning - ICML 2004, p. 18 (2004)
7. Daszykowski, M., Walczak, B.: Density-based clustering methods. Compr. Chemom. **2**, 635–654 (2010). https://doi.org/10.1016/B978-044452701-1.00067-3
8. Fern, X.Z., Lin, W.: Cluster ensemble selection. Stat. Anal. Data Min. **1**(3), 128–141 (2008). https://doi.org/10.1002/sam.10008
9. Ferrari, D.G., De Castro, L.N.: Clustering algorithm selection by meta-learning systems: a new distance-based problem characterization and ranking combination methods. Inf. Sci. **301**, 181–194 (2015)
10. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics **21**(3), 768–769 (1965)
11. Ghosh, J., Acharya, A.: Wiley interdisciplinary reviews: data mining and knowledge discovery. Clust. Ensembles **1**(4), 305–315 (2011)
12. Horst, R., Thoai, N.V.: Dc programming: overview. J. Optim. Theor. Appl. **103**(1), 1–43 (1999)
13. Jia, J., Xiao, X., Liu, B., Jiao, L.: Bagging-based spectral clustering ensemble selection. Pattern Recognit. Lett. **32**(Fskd), 1456–1467 (2011)

14. Le Thi, H.A., Pham Dinh, T., Le, H.M., Vo, X.T.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**, 26–46 (2015). https://doi.org/10.1016/j.ejor.2014.11.031
15. Mangasarian, O.: Machine learning via polyhedral concave minimization. In: Applied Mathematics and Parallel Computing, pp. 175–188. Springer, Heidelberg (1996)
16. Moon, T.: The expectation-maximization algorithm. IEEE Sig. Process. Mag. **13**(6), 47–60 (1996)
17. Muñoz, A., Diego, I.: From indefinite to positive semi-definite matrices. Struct. Syntactic Stat. Pattern Recognit. **4109**, 764–772 (2006)
18. Pham Dinh, T., Le Thi, H.A.: Recent advances in DC programming and DCA. In: Nguyen, N.T., Le-Thi, H.A. (eds.) Trans. Comput. Intell. XIII, pp. 1–37. Springer, Heidelberg (2014)
19. Rinaldi, F., Schoen, F., Sciandrone, M.: Concave programming for minimizing the zero-norm over polyhedral sets. Comput. Optim. Appl. (2010)
20. Rockafeller, R.T.: Convex analysis (1970)
21. Schapire, R.: The boosting approach to machine learning: an overview (2001)
22. Shen, X., Diamond, S., Gu, Y., Boyd, S.: Disciplined convex-concave programming. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 1009–1014. IEEE (2016)
23. Sriperumbudur, B.K., Torres, D.A., Lanckriet, G.R.: A majorization-minimization approach to the sparse generalized eigenvalue problem. Mach. Learn. (2011)
24. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
25. Tao, P.D., et al.: Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In: North-Holland Mathematics Studies, vol. 129, pp. 249–271. Elsevier (1986)
26. Thiao, M., Dinh, T.P., Le Thi, H.A.: Dc programming approach for a class of nonconvex programs involving l 0 norm. In: Modelling, Computation and Optimization in Information Systems and Management Sciences, pp. 348–357. Springer (2008)
27. Tibshirani, R.: Regression selection and shrinkage via the lasso (1996). https://doi.org/10.2307/2346178
28. Weston, J., Elisseeff, A., Scholkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. (2003)
29. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)
30. Yu, Z., Li, L., Gao, Y., You, J., Liu, J., Wong, H.S., Han, G.: Hybrid clustering solution selection strategy. Pattern Recognit. **47**(10), 3362–3375 (2014)
31. Zhang, Y., Burer, S., Street, W.N.: Ensemble pruning via semi-definite programming. J. Mach. Learn. Res. **7**, 1315–1338 (2006)