Difference Of Convex (DC) Functions and DC Programming

Songcan Chen

Outline

1. A Brief History
2. DC Functions and their Property
3. Some examples
4. DC Programming
5. Case Study
6. Our next work

1. A Brief History

- 1964, Hoang Tuy, (incidentally in his convex optimization paper),
- 1979, J. F. Toland, Duality formulation
- 1985, Pham Dinh Tao, DC Algorithm
- 1990 --, Pham Dinh Tao, et al
- …

H. Tuy, *Concave programming under linear constraints,* Translated Soviet Mathematics 5 (1964), 1437-1440.
J. F. Toland, *On subdi®erential calculus and duality in nonconvex optimization*, Bull. Soc. Math. France, M¶emoire 60 (1979), 173-180.
Pham Dinh Tao, *Duality in d.c. (di®erence of convex functions) optimization. Subgradient methods*, Trends in Mathematical Optimization, International Series of Numer Math. 84 (1988), Birkhauser, 277-293.

Applicable Fields

- For smooth/non-smooth and convex/non-convex optimization problems, especially,
- For large-scale DC problems → Robust and efficient in solving!

Hence

- Machine learning (Clustering, Kernel optimization, Feature selection,…)
- Engineering (Quality control,...)
- …

2.1 DC Functions

- **Definition 2.1.** Let *C* be a convex subset of R^n. A real-valued function *f* : C → R is called DC on C, if there exist two convex functions *g, h:* C → R such that *f* can be expressed in the form

$$f(x)=g(x)-h(x) \tag{1}$$

h(x) convex → -h(x) concave.

If C= R^n, then f is simply called a DC function.

Notice: DC representation for f is NOT unique, in fact, can have infinite decompositions!

2.2 Their Properties

Let f and f_i, i = 1 , . . . , *m,* be DC functions. Then, the following functions are also DC:

1) $\sum_{i=1}^{m} \lambda_i f_i, \qquad \lambda_i \in R, i = 1, 2, ..., m.$

2) $\max_{i=1,2,...,m} \{f_i\}$ *and* $\min_{i=1,2,...,m} \{f_i\}$

3) $|f(x)|$

4) $\prod_{i=1}^{m} f_i$

2.2 Their Properties (Cont'd)

1) Every function f: $R^n \rightarrow R$ whose second partial derivatives are continuous everywhere is DC.
2) Let *C* be a compact convex subset of R^n. Then for any continuous function $c: C \rightarrow R$ and for any $\varepsilon > 0$, there exists a DC function f: $C \rightarrow R$ such that

$$|c(x) - f(x)| < \varepsilon, \text{ for any } x \text{ in } C.$$

3) Let f: $R^n \rightarrow R$ be DC, and let g: $R \rightarrow R$ be convex. Then, the composite function $(g \circ f)(x) = g(f(x))$ is DC.

3. Some simple examples

1) $x^t Qx$, $Q=A-B$, A and B are positive semi-definite.
2) $x^t y$,
3) Let d_M be a distance function, then

   $d_M(x)=\inf\{||x-y||: y \text{ in } M\}$.

Proof of 3)

Proof: We have

$$
\begin{aligned}
d_M^2(x) &= \inf\{||x-y||^2 : y \in M\} \\
&= ||x||^2 + \inf\{-||x||^2 + ||x-y||^2 : y \in M\} \\
&= ||x||^2 - \sup\{||x||^2 - ||x-y||^2 : y \in M\} \\
&= ||x||^2 - \sup\{2x^T y - ||y||^2 : y \in M\}.
\end{aligned}
$$

The norm $p(x) = ||x||^2$ is convex, and the function $q(x) := \sup\{2x^T y - ||y||^2 : y \in M\}$ is the pointwise supremum of a family of affine functions, and hence convex.

4. DC Programming

- 4.1 Primal Problem
- 4.2 Dual Problem
- 4.3 DC Algorithm (DCA)

4.1 Primal Problem

- A general form

$$(P_{dc}) \quad \alpha = \inf\{f_0(x) : x \in X \subseteq R^n, f_i(x) \le 0, i = 1, 2, \ldots, m\}$$

Where $f_i=g_i-h_i$, $i=1,2,\ldots,m$ are DC functions and X is a closed convex subset of R^n.

Constrained (closed) Set X can be represented by a convex indicator function which is added to the $g_0(x)$ ($f_0=g_0-h_0$): $I_X(x)=0$ if x in X, $+\infty$ otherwise.

4.1 Primal Problem (Cont'd)

When X is constrained by a set of linear inequality equations and the objective function is linear, the optimization problem is called polyhedral DC, solving it amounts to solving a linear programming.

For example, selecting features based on SVM and l_0 norm [5].

Notations

1) The conjugate function *g** of *g is* defined by

$$g^*(y) = \sup\{\langle x, y\rangle - g(x) : \; x \in X\}$$

2) Support Domain of g(x)

$$\mathrm{dom}\; g = \{x \in X : \; g(x) < +\infty\}$$

3) ε -subdifferential of g(x) at x^0, when ε =0, simply called subdifferential.

$$\partial_\epsilon g(x^o) = \{y \in Y : g(x) \geq g(x^o) + \langle x - x^o, y\rangle - \epsilon \quad \forall x \in X\}$$

Notations (Cont'd)

Support Domain of the subdifferential ∂g

$$\mathrm{dom}\ \partial g = \{x \in X : \ \partial g(x) \neq \emptyset\}$$

Range Domain of ∂g

$$\mathrm{range}\, \partial g = \cup \{\partial g(x) : \ x \in \mathrm{dom}\ \partial g\}$$

4.2 Dual Problem

Using the definition of conjugate functions, we have

$$\begin{aligned}\alpha &= \inf\{g(x) - h(x) : x \in X\} \\ &= \inf\{g(x) - \sup\{\langle x, y\rangle - h^*(y) : y \in Y\} : x \in X\} \\ &= \inf\{\beta(y) : \ y \in Y\}\end{aligned}$$

with

$$(P_y) \qquad \beta(y) = \inf\{g(x) - (\langle x, y\rangle - h^*(y)) \ : x \in X\}$$

$$\Rightarrow \quad \beta(y) = h^*(y) - g^*(y) \text{ if } y \in \operatorname{dom} h^*, \ +\infty \text{ otherwise}$$

4.2 Dual Problem (Cont'd)

Dual Formulation:

$$(\mathrm{D}) \qquad \alpha = \inf\{h^*(y) - g^*(y) : y \in Y\}$$

Where Y= dom ∂h*.

A perfect symmetry exists between the primal and its dual programs (P) and (D):

the dual program to (D) is exactly (P).

4.2 Dual Problem (Cont'd)

- The necessary local optimality condition for P_{dc}, is

$$\partial h(x^*) \text{ in } \partial g(x^*)$$

- A point that x^* that verifies the generalized Kuhn-Tucker condition

$$\partial h(x^*) \cap \partial g(x^*) \neq \varnothing$$

is called a critical point of $g-h$.

4.3 DCA

DCA Scheme

INPUT

– Let $x^0 \in \mathbb{R}^p$ be a best guest, $0 \leftarrow k$.

REPEAT

– Calculate $y^k \in \partial h(x^k)$.
– Calculate

$$x^{k+1} \in \arg\min \left\{ g(x) - h(x^k) - \langle x - x^k, y^k \rangle \quad s.t. x \in \mathbb{R}^p \right\}. \qquad (P_k)$$

– $k + 1 \leftarrow k$.

UNTIL{convergence of x^k.}

Affine majorization of the concave part –h(x)!

4.3 DCA (Cont’d)

- Different decompositions → thus make trade-off between Complexity of each step,
  - number of iterations.
  - Local convergence, empirically: “good” optima.

4.3 DCA (Cont'd)

Convergence properties

– DCA is a descent method (i.e., the sequences $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are both decreasing) *without linesearch*;

– If the optimal value α of problem (P_{dc}) is finite and the infinite sequences $\{x^k\}$ and $\{y^k\}$ are bounded, then every limit point x^* (resp. y^*) of $\{x^k\}$ (resp. $\{y^k\}$) is a critical point of $g - h$ (resp. $h^* - g^*$), i.e., $\partial h(x^*) \cap \partial g(x^*) = \varnothing$ (resp. $\partial h^*(y^*) \cap \partial g^*(y^*) = \varnothing$).

– DCA has a *linear convergence* for general DC programs.

5. Case Study

- 5.1 Fuzzy c-means Clustering
- 5.2 Feature Selection and Classification

5.1 Fuzzy c-means Clustering

$$\begin{cases} \min J_m(U,V) := \sum_{k=1}^{n} \sum_{i=1}^{c} u_{i,k}^m ||x_k - v_i||^2 \\ s.t \quad u_{i,k} \in [0,1] \text{ for } i = 1,..,c \quad k = 1,..,n \\ \sum_{i=1}^{c} u_{i,k} = 1, \; k = 1,..,n \end{cases}$$

FCM (Cont'd)

- How to be changed to DC

  1) g and h?

  2) X – a convex set of variables (U, V)?

Characterization of Convex Set

- From the centers' solution $V=\{v_i, \ i=1,2,\ldots,c\}$,

$$v_i \sum_{k=1}^{n} u_{i,k}^{m} = \sum_{k=1}^{n} u_{i,k}^{m} x_k$$

$$\Rightarrow$$

$$\|v_i\|^2 \le \frac{\left(\sum_{k=1}^{n} u_{i,k}^{m} \|x_k\|\right)^2}{\left(\sum_{k=1}^{n} u_{i,k}^{m}\right)^2} \le \sum_{k=1}^{n} \|x_k\|^2 := r^2$$

Leading to the Euclidean ball R_i with radius r. It is convex!

In fact, $\| v_i \| \leqq \max\{ \| x_k \|, k=1, 2, \ldots, n\}$ for all i.

Characterization of Convex Set

- For U, let $u_{i,k} = t_{i,k}^2$
  - Constraints

$$\sum_{i=1}^{c} u_{i,k} = 1$$

$$\Rightarrow \quad \sum_{i=1}^{c} t_{i,k}^2 = 1 \text{ or } \|t_k\|^2 = 1 \text{ with } t_k \in \mathbb{R}^c$$

Leading to the Euclidean sphere S_k with radius 1. It is NOT convex.

Equivalent Formulation to FCM

$$\begin{cases} \min J_{2m}(T,V) := \sum_{k=1}^{n} \sum_{i=1}^{c} t_{i,k}^{2m} ||x_k - v_i||^2 \\ s.t \quad T \in \mathcal{S} := \Pi_{k=1}^{n} \ S_k, \ V \in \mathcal{C} := \Pi_{i=1}^{c} R_i \end{cases}$$

A DC decomposition of the above objective function

$$\begin{aligned} J_{2m}(T,V) = \frac{\rho}{2}(\|T\|^2 + \|V\|^2) \\ - \left[\frac{\rho}{2}\|(T,V)\|^2 - J_{2m}(T,V)\right] \end{aligned}$$

DC Formulation

For all $(T,V) \in \mathcal{S} \times \mathcal{C}$

$$J_{2m}(T,V) = \frac{\rho}{2}n + \frac{\rho}{2}\|V\|^2 - H(T,V)$$

with $H(T,V) := \frac{\rho}{2}\|(T,V)\|^2 - J_{2m}(T,V)$

A Question: is H(T, V) unconditionally convex? No!

Condition ensuring H(T,V)

Proposition 1. *Let $\mathcal{B} := \Pi_{k=1}^{n} B_k$, where B_k is the ball of centre* 0 *and radius* 1 *in* $\mathbb{R}^c$. *The function $H(T,V)$ is convex on $\mathcal{B} \times \mathcal{C}$ for all values of ρ such that*

$$\rho \geq \frac{m}{n}(2m-1)\alpha^2 + 1 + \sqrt{\left[\frac{m}{n}(2m-1)\alpha^2 + 1\right]^2 + \frac{16}{n}m^2\alpha^2}, \tag{8}$$

where

$$\alpha = r + \max_{1 \leq k \leq n} \|x_k\| . \tag{9}$$

Notice here B denotes a Ball and thus is convex!

In fact, *alpha* can be 2 max{ ‖ xk ‖ , k=1, 2, …, n} !

Proof (1)

Proof: from
$$H(T,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left[\frac{\rho}{2}t_{i,k}^2 + \frac{\rho}{2n}\|v_i\|^2 - t_{i,k}^{2m}\|x_k - v_i\|^2\right]$$

Just prove the function are convex for all *i* and *k*

$$h_{i,k}(t_{i,k}, v_i) := \frac{\rho}{2}t_{i,k}^2 + \frac{\rho}{2n}\|v_i\|^2 - t_{i,k}^{2m}\|x_k - v_i\|^2$$

Define
$$f: \quad \mathbb{R}\times\mathbb{R}\to\mathbb{R}$$
$$f(x,y) = \tfrac{\rho}{2}x^2 + \tfrac{\rho}{2n}y^2 - x^{2m}y^2$$

Its Hessian

$$J_f(x,y) = \begin{pmatrix} \rho - 2m(2m-1)y^2x^{2m-2} & -4mx^{2m-1}y \\ -4mx^{2m-1}y & \frac{\rho}{n} - 2x^{2m} \end{pmatrix}$$

Proof (2)

For all (x, y): $0 \leq x \leq 1; ||y|| \leq \alpha$

$$
\begin{aligned}
\mid J_f(x,y) \mid &= \left(\rho - 2m(2m-1)y^2x^{2m-2}\right)\left(\frac{\rho}{n} - 2x^{2m}\right) - 16m^2x^{4m-2}y^2 \\
&\geq \frac{1}{n}\rho^2 - \left[2\frac{m}{n}(2m-1)y^2x^{2m-2} + 2x^{2m}\right]\rho - 16m^2x^{4m-2}y^2 \\
&\geq \frac{1}{n}\rho^2 - 2\left(\frac{m}{n}(2m-1)\alpha^2 + 1\right)\rho - 16m^2\alpha^2.
\end{aligned}
$$

So f(x, y) is convex on $[0,1] \times [-\alpha, \alpha]$

Proof (3)

implying

$$\theta_{i,k}(t_{i,k}, v_i) := \frac{\rho}{2}\, t_{i,k}^2 + \frac{\rho}{2n}\|x_k - v_i\|^2 \;\; - t_{i,k}^{2m}\|x_k - v_i\|^2$$

is convex on $\{0 \le t_{i,k} \le 1, \|v_i\| \le r\}$

Further $h_{i,k}$ is convex

$$h_{i,k}(t_{i,k}, v_i) = \theta_{i,k}(t_{i,k}, v_i) + \frac{\rho}{n}\langle x_k, v_i\rangle - \frac{\rho}{2n}\|x_k\|^2$$

Finally, the function $H(T,V)$ is convex on $B \times C$.

Proof (4)

For all $T \in B$ (*closed ball*) and a given matrix $V \in C$, the function $J_{2m}(T,V)$ is concave in variable T (since $H(T,V)$ is convex). Hence S (*sphere, i.e., boundary*) contains minimizers (reaching at boundary) of $J_{2m}(T,V)$ on B, i.e.,

$$\min\left\{\frac{\rho}{2}\|V\|^2 - H(T,V) : (T,V) \in \mathcal{B} \times \mathcal{C}\right\}$$

$$= \min\left\{\frac{\rho}{2}\|V\|^2 - H(T,V) : (T,V) \in \mathcal{S} \times \mathcal{C}\right\}$$

DC Formulation

$$\min\left\{\frac{\rho}{2}\|V\|^2 - H(T,V) : \ (T,V) \in \mathcal{B}\times\mathcal{C}\right\}$$

$$\min\begin{cases}\chi_{\mathcal{B}\times\mathcal{C}}(T,V) + \frac{\rho}{2}\|V\|^2 - H(T,V)\\ s.t.\ (T,V) \in \mathbb{R}^{c\times n}\times\mathbb{R}^{c\times p}.\end{cases}$$

$$\chi_{\mathcal{B}\times\mathcal{C}}(T,V) + \frac{\rho}{2}\|V\|^2 - H(T,V) := G(T,V) - H(T,V)$$

where $G(T,V) := \chi_{\mathcal{B}\times\mathcal{C}}(T,V) + \frac{\rho}{2}\|V\|^2$

Solving FCM by DCA (1)

A key: construct two sequences $(Y^l, Z^l) \in \partial H(T^l, V^l)$ and

$$(T^{l+1}, V^{l+1}) \in \arg\min \begin{cases} \frac{\rho}{2} \|V\|^2 - \langle (T, V), (Y^l, Z^l) \rangle \\ s.t. \ (T, V) \in \mathcal{B} \times \mathcal{C}. \end{cases}$$

H is differentiable and its gradient at the point (T^l, V^l):

$$\nabla H(T^l, V^l) = \rho(T^l, V^l) - \left(2m t_{i,k}^{2m-1} \|x_k - v_i\|^2, 2 \sum_{k=1}^{n} (v_i - x_k) t_{i,k}^{2m}\right) \qquad (14)$$

Algorithm 1. DCA applied to FCM

INPUT

- $T^0 \in \mathbb{R}^{c \times n}$ and $V^0 \in \mathbb{R}^{c \times p}$.
- $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

REPEAT

- Calculate $(Y^l, Z^l) = \nabla H(T^l, V^l)$ via (14);
- Calculate (T^{l+1}, V^{l+1}) via (15) and (16);
- $l + 1 \leftarrow l$.

UNTIL$\{\|(T^{l+1}, V^{l+1}) - (T^l, V^l)\| \leq \epsilon(\|(T^{l+1}, V^{l+1})\|)\}$

Solving FCM by DCA (2)

$$T^{l+1} = \mathrm{Proj}_{\mathcal{B}}(Y^l), V^{l+1} = \mathrm{Proj}_{\mathcal{C}}(\frac{1}{\rho}Z^l)$$

More precisely:

$$V_{i,.}^{l+1} = \begin{cases} \frac{(Z^l)_{i,.}}{\rho} \text{ if } \|(Z^l)_{i,.}\| \leq \rho r \\ \frac{(Z^l)_{i,.}r}{\|(Z^l)_{i,.}\|} \text{ otherwise} \end{cases}, i = 1,..,c, \tag{15}$$

$$T_{.,k}^{l+1} = \begin{cases} Y_{.,k}^l \text{ if } \|Y_{.,k}^l\| \leq 1 \\ \frac{(Y^l)_{.,k}}{\|(Y^l)_{.,k}\|} \text{ otherwise} \end{cases}, k = 1,..,n. \tag{16}$$

Accelerating DCA -- FCM-DCM (1)

Algorithm 2. Combined FCM-DCA algorithm

INPUT

- Let U^0 and V^0 be the membership and the cluster centers randomly generated.
- Set $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

REPEAT

i. One iteration of FCM:

Accelerating DCA -- FCM-DCM (2)

– Compute the cluster centers V^l via

$$v_i = \sum_{k=1}^{n} u_{ik}^m x_k / \sum_{k=1}^{n} u_{ik}^m \quad \forall i = 1, .., c. \tag{17}$$

– Compute the membership U^l via

$$u_{ik} = \left[\sum_{j=1}^{c} \frac{\|x_k - v_i\|^{2/(m-1)}}{\|x_k - v_j\|^{2/(m-1)}} \right]^{-1}. \tag{18}$$

– Set $t_{ik} = \sqrt{u_{ik}}$, $\forall i = 1, .., c$ and $\forall k = 1, .., n$.

Accelerating DCA -- FCM-DCM (3)

ii. One iteration of DCA:

- Calculate $(Y^l, Z^l) = \nabla H(T^l, V^l)$ via (14);
- Calculate (T^{l+1}, V^{l+1}) via (15) and (16);
- $l + 1 \leftarrow l$

UNTIL$\{\|(T^{l+1}, V^{l+1}) - (T^l, V^l)\| \leq \epsilon(\|(T^{l+1}, V^{l+1})\|)\}$

Two phase algorithm 3

INPUT

– Let U^0 and V^0 be the membership and the cluster centers randomly generated.
– Set $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

PHASE 1:

– Perform q iterations of **Algorithm 2** for obtaining (T^{q+1}, V^{q+1}).
– Update $(T^0, V^0) \leftarrow (T^{q+1}, V^{q+1})$

PHASE 2:

– Apply **Algorithm 1** from the initial point (T^0, V^0) until the convergence.

Partial results

Table 1. Computation time of **FCM Algorithm** and **Algorithm 2, 3**

Data			FCM		Algorithm 2		Algorithm 3		
$N°$	Size	c	N^oF	Time	N^oI	Time	q	N^oD	Time
1	128^2	2	24	1.453	16	1.312	12	10	1.219
2	128^2	2	17	1.003	12	0.985	10	2	0.765
3	256^2	3	36	15.340	24	13.297	20	2	10.176
4	256^2	3	75	31.281	57	30.843	30	12	26.915
5	256^2	3	39	15.750	27	14.687	20	14	13.125
6	256^2	5	91	84.969	75	86.969	40	78	61.500
7	256^2	3	73	31.094	62	34.286	15	21	24.188
8	256^2	3	78	34.512	52	32.162	20	13	29.182
9	512^2	3	49	92.076	41	102.589	30	46	74.586
10	512^2	5	246	915.095	196	897.043	120	86	691.854

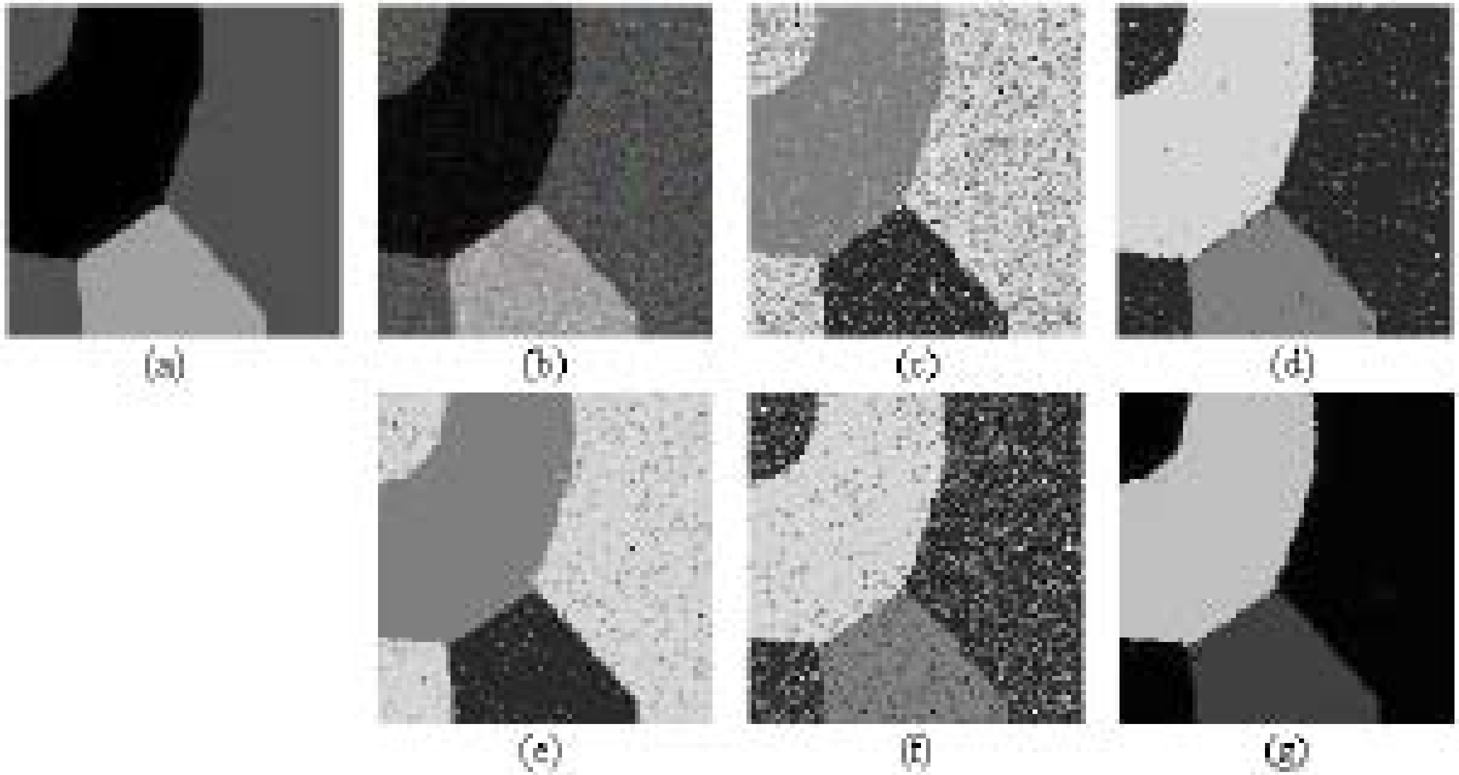


Fig. 1. The original noisy image and the results of segmentation (c=3)

(a) (resp. (b)) corresponds to the original image without (resp. with) noise;

(c) (resp. (d)) represents the resulting image given by FCM Algorithm without (resp. with) spatial information.

(e) represents the resulting image given by **Algorithm 2** without spatial information

(f) (resp. (g)) represents the resulting image given by **Algorithm 3** without (resp. with) spatial information.

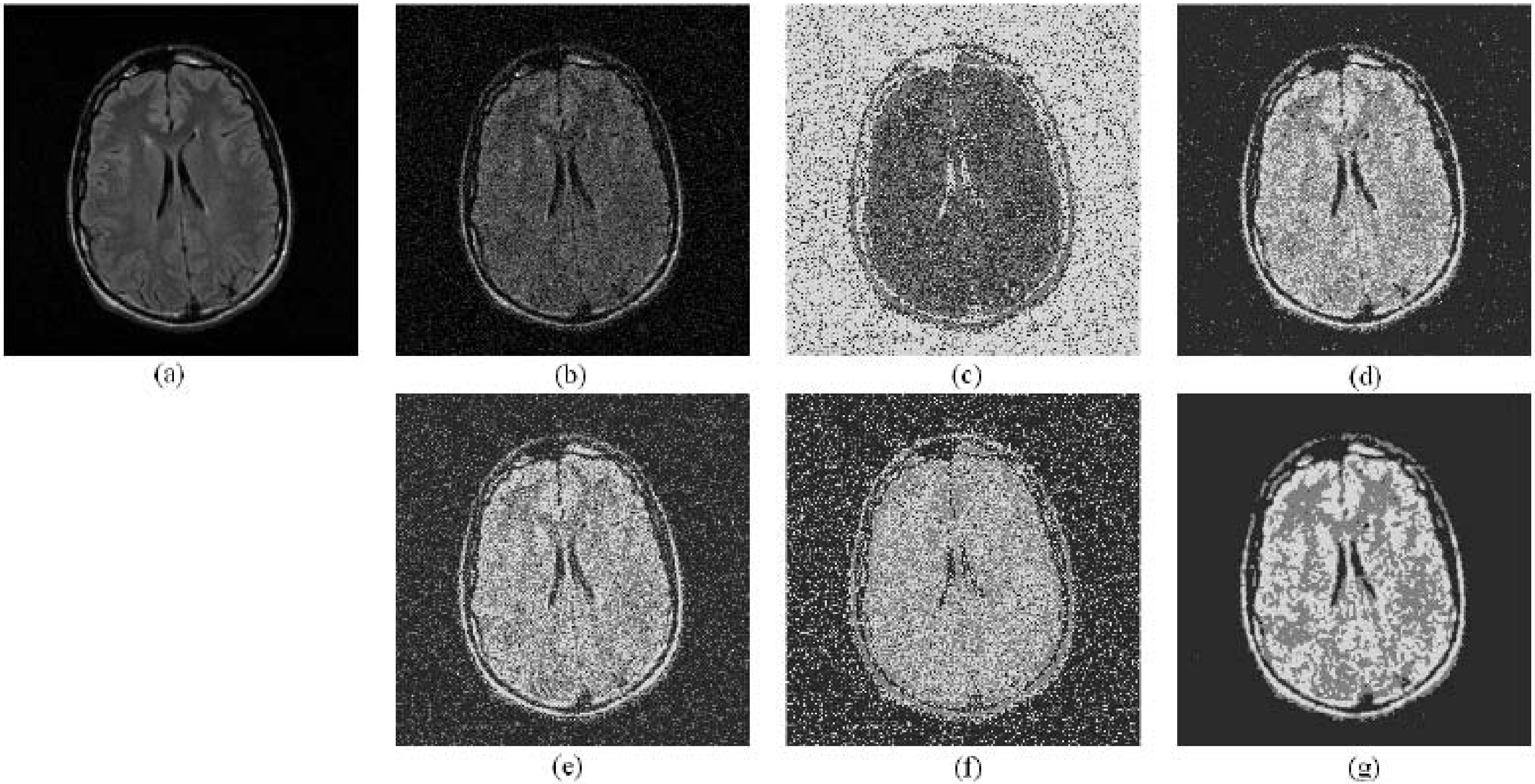


Fig. 2. The medical noisy image and the results of segmentation (c=3)

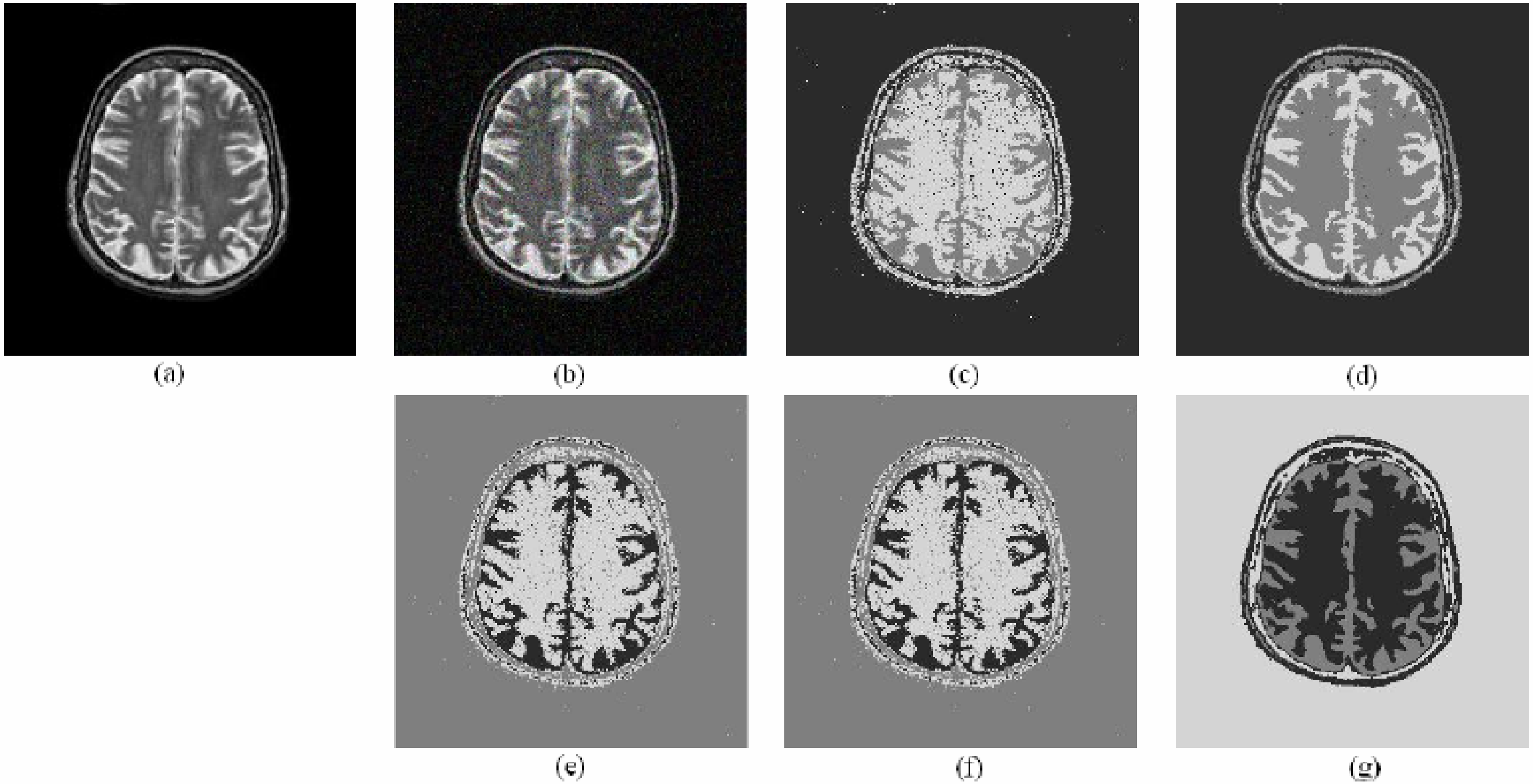


Fig. 3. The medical noisy image and the results of segmentation (c=3)

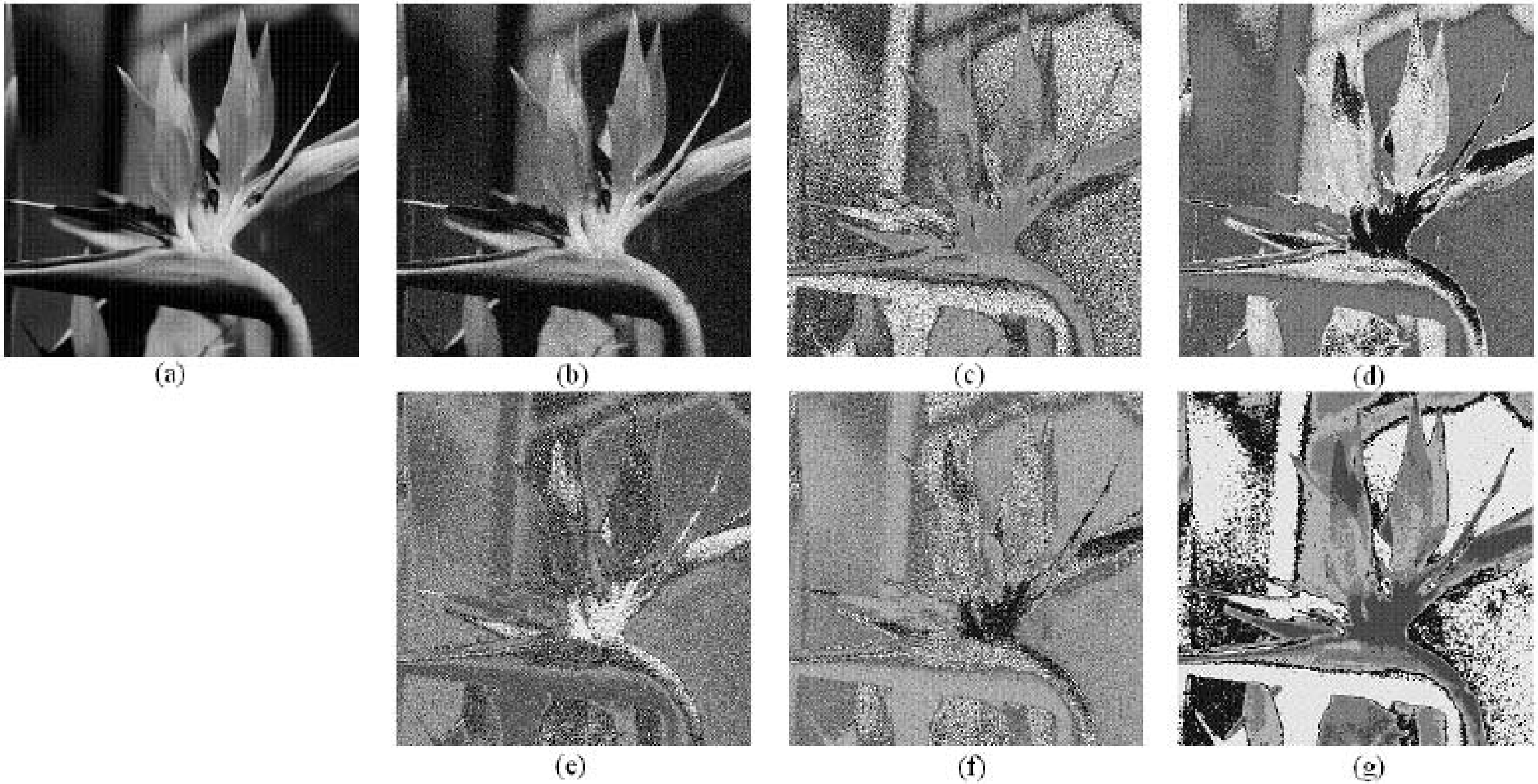


Fig. 4. The **Blume** noisy image and the results of segmentation (c=5)

5.2 Feature Selection and Classification

Formulation of problem

- Given two finite point sets A and B in R^n represented by the matrices $A \in R^{m \times n}$ and $B \in R^{k \times n}$, respectively. Discriminate these sets by a separating plane ($w \in R^n$, $\gamma \in R$)

$$P = \{x \mid x \in \mathbb{R}^n, x^T w = \gamma\} \qquad (1)$$

which uses as few features as possible.

The optimization problem

$$
\begin{array}{ll}
\min_{w,\gamma,y,z} & (1-\lambda)(\frac{1}{m}e^T y + \frac{1}{k}e^T z) + \lambda \|w\|_0 \\
s.t & -Aw + e\gamma + e \le y \\
 & Bw - e\gamma + e \le z \\
 & y \ge 0,\ z \ge 0.
\end{array} \quad (2)
$$

Where y_i, i=1,2,…,m and z_j, j=1,2,…,k are non-negative slack variables, e is a vector with all entries of 1.

The zero-norm: $\|w\|_0 := card\{w_i : w_i \neq 0\}$

P. S. Bredley and O. L. Mangasarian, *Feature Selection via concave minimization and support vector machines,* ICML'08.

Optimization Difficulty of Zero-Norm

- Discontinuity at the origin
- NP-Hard

Solution: Approximation to Zero-norm!

for example,

$$\|v\|_0 \simeq e^T\left(e - \varepsilon^{-\alpha v}\right).$$

Approximate Zero-norm

$$\|w\|_0 \simeq \sum_{i=1}^{n} \eta(\alpha, w_i).$$

where

$$\eta(x, \alpha) = \begin{cases} 1 - \varepsilon^{-\alpha x} & \text{if } x \geq 0 \\ 1 - \varepsilon^{\alpha x} & \text{if } x < 0 \end{cases} \quad , \alpha > 0.$$

Reformulation of the optimization

$$\min \left\{ \begin{array}{l} F(y,z,w,\gamma) := (1-\lambda)(\frac{e^T y}{m} + \frac{e^T z}{k}) \\ +\lambda \sum_{i=1}^{n} \eta(w_i) : (y,z,w,\gamma) \in K \end{array} \right\}$$

where K is the polyhedral convex set defined by:

$$K := \left\{ \begin{array}{l} (y,z,w,\gamma) \in \mathbb{R}^{m+k+n+1} : \\ -Aw + e\gamma + e \leq y, \\ Bw - e\gamma + e \leq z \end{array} \right\}.$$

A DC decomposition of the approximation

$$\eta(x) = g(x) - h(x)$$

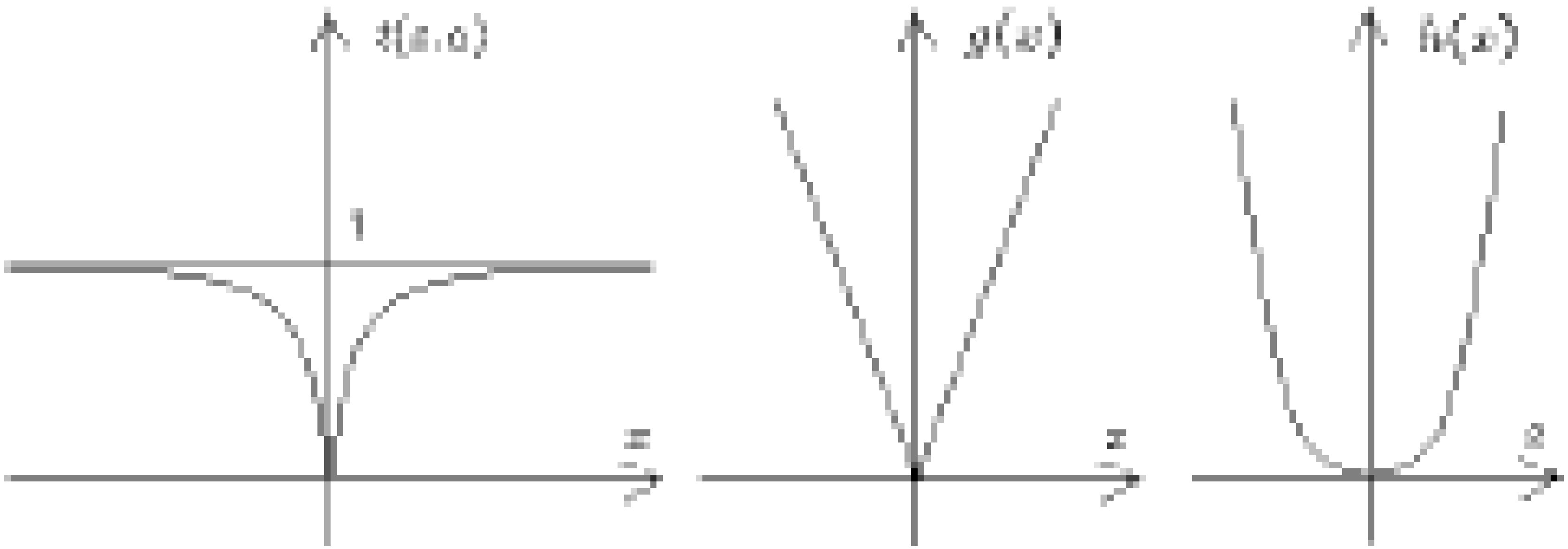
where

$$g(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ -\alpha x & \text{if } x < 0 \end{cases}$$

$$h(x) = g(x) - \eta(x) = \begin{cases} \alpha x - 1 + \varepsilon^{-\alpha x} & \text{if } x \geq 0 \\ -\alpha x - 1 + \varepsilon^{\alpha x} & \text{if } x < 0 \end{cases}$$

They are both convex!

Illustration



DC Decomposition of the Objective

$$F(y,z,w) := G(y,z,w) - H(y,z,w)$$

where:

$$G(y,z,w) := (1-\lambda)(\frac{e^T y}{m} + \frac{e^T z}{k}) + \lambda \sum_{j=1}^{n} g(w_j)$$

$$H(y,z,w) := \lambda \sum_{j=1}^{n} h(w_j)$$

A Final Formulation: FSC-DC

$$\min \quad \{G(y,z,w) - H(y,z,w) : (y,z,w,\gamma) \in K\}$$

$$\Longleftrightarrow$$

$$\min \left\{ \begin{array}{l} \chi_K(y,z,w,\gamma) + G(y,z,w) - H(y,z,w) : \\ (y,z,w,\gamma) \in \mathbb{R}^m \times \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}, \end{array} \right\} \tag{13}$$

DCA Revisited

Generic DCA scheme:
Initialization: Let $x^0 \in \mathbb{R}^p$ be a best guest,
$0 \leftarrow k$.
iteration k = 0, 1, ...
Calculate $y^k \in \partial H(x^k)$
Calculate

$$x^{k+1} \in \arg\min \left\{ \begin{array}{l} G(x) - H(x^k) - \langle x - x^k, y^k \rangle : \\ x \in \mathbb{R}^p \end{array} \right\}$$

$k + 1 \leftarrow k$
Until convergence of x^k.

DCA for FSC

Initialization Let τ be a tolerance sufficiently small, set $k = 0$.

Choose $(y^0, z^0, w^0, \gamma^0) \in \mathbb{R}^m \times \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}$.

Repeat

- Compute $v^k \in \partial H(w^k)$ via (15).
- Solve the linear program (16) to obtain $(y^{k+1}, z^{k+1}, w^{k+1}, \gamma^{k+1})$
- $k + 1 \leftarrow k$

Until

$$\left\|y^k - y^{k-1}\right\| + \left\|z^k - z^{k-1}\right\| + \left\|w^k - w^{k-1}\right\| + \left|\gamma^k - \gamma^{k-1}\right| \leq \tau\left(1 + \left\|y^k\right\| + \left\|z^k\right\| + \left\|w^k\right\| + \left|\gamma^k\right|\right)$$

(15) And (16)

$$v_j = \begin{cases} \alpha(1-\varepsilon^{-\alpha w_j}) & \text{if } w_j \geq 0 \\ -\alpha(1-\varepsilon^{\alpha w_j}) & \text{if } w_j < 0 \end{cases} \quad . \tag{15}$$

$$\min\{G(y,z,w) - \langle v^k, w\rangle : \quad (y,z,w,\gamma) \in K\}$$

$$= \min \left\{ \begin{array}{c} (1-\lambda)(\frac{e^T y}{m} + \frac{e^T z}{k}) + \\ \lambda \sum_{j=1}^{n} \max\{\alpha w_j, -\alpha w_j\} - \langle v^k, w\rangle \\ \text{s.t.} \quad (y,z,w) \in K \end{array} \right\}$$

$$\Leftrightarrow \min \left\{ \begin{array}{l} (1-\lambda)(\frac{e^T y}{m} + \frac{e^T z}{k}) + \lambda \sum_{i=1}^{n} t_j \\ -\langle v^k, w\rangle : (y,z,w,\gamma,t) \in \Omega \end{array} \right\} \tag{16}$$

Feasible Domain

$$\Omega := \left\{ \begin{array}{l} (y, z, w, \gamma, t) \in \mathbb{R}^{m+k+n+1+n} : \\ (y, z, w, \gamma) \in K, \\ -\alpha w_j \leq t_j, \alpha w_j \leq t_j, j = 1..n \end{array} \right\}$$

An Important Theorem

Theorem 1 *(Convergence properties of Algorithm DCA)*

(i) *DCA generates a sequence* $\{(y^k, z^k, w^k, \gamma^k)\}$ *such that the sequence* $\{F(y^k, z^k, w^k)\}$ *is monotonously decreasing.*

(ii) *The sequence* $\{(y^k, z^k, w^k, \gamma^k)\}$ *converges to* $(y^*, z^*, w^*, \gamma^*)$ *after a finite number of iterations.*

(iii) *The point* (y^*, z^*, w^*) *is a critical point of the objective function* F *in Problem (13).*

Experimental Result

Data set	FSV			DCA		
	selected feature (%)	correctness (%)		selected feature (%)	correctness (%)	
		train	test		train	test
Pima Indian	66	75.22	**74.60**	50	**76.02**	71.18
BUPA Liver	75	68.18	65.20	50	**87.44**	**85.99**
Ionosphere	31	**90.47**	**84.07**	9	73.50	63.24
WPBC (24 mo)	12	73.97	66.42	9	**80.00**	**75.13**
WPBC (60 mo)	8	70.70	67.05	6	**74.40**	**72.50**
Average	38.4	75.70	71.47	**24.8**	**78.36**	**73.42**

Selection of the λ: CV

Step 1. Set aside 10% of the training data as a "tuning" set.

Step 2. Obtain a classifier for the given value of λ.

Step 3. Determine correctness on the "tuning" set.

Step 4. Repeat steps 1-3 10 times, each time setting aside a different 10% portion of the training data. The "score" for this value of λ is the average of the 10 correctness values determined in Step 3.

6. Our next work: Applying DCP

- (Structured) AUC-FSC-DC (based SVM)
- Asymmetric (SVM) FSC-DC
- FSC based on DR and LR
- KFCM-DC and Combination
- NMF-DC and Combination
- SPP-DC (mainly in Sparse solving)
- …

Reference

[1] R. Horst, N. V. THOAI, DC Programming: Overview, JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 103, No. 1, pp. 1-43, 1999.

[2] PHAM DINH TAO AND LE THI HOAI AN, CONVEX ANALYSIS APPROACH TO D.C. PROGRAMMING, ACTA MATHEMATICA VIETNAMICA, 22(1) 1997, pp. 289-355.

[3] Le Thi Hoai An, D.C. Programming for Solving a Class of Global Optimization Problems via Reformulation by Exact Penalty, C. Bliek et al. (Eds.): COCOS 2002, LNCS 2861, pp. 87–101, 2003.

[4] Julia Neumann, Christoph Schnorr, and Gabriele Steidl, SVM-Based Feature Selection by Direct Objective Minimisation, C.E. Rasmussen et al. (Eds.): DAGM 2004, LNCS 3175, pp. 212–219, 2004.

[5] Hoai An Le Thi, Hoai Minh Le, Van Vinh Nguyen, Tao Pham Dinh, A DC programming approach for feature selection in support vector machines learning, Adv Data Anal Classif (2008) 2:259–278.

[6] Le Thi Hoai An, M. Taye Belghiti, Pham Dinh Tao, Feature Selection via DC Programming and DCA, J Glob Optim (2007) 37:593–608.

[7] Hoai An Le Thi, Hoai Minh Le, Tao Pham Dinh, Fuzzy clustering based on nonconvex optimisation approaches using difference of convex (DC) functions algorithms, ADAC (2007) 1:85–104.

[8] Le ThiHoai An, M. Tayeb Belghiti, Pham Dinh Tao, A new efficient algorithm based on DC programming and DCA for Clustering, J Glob Optim (2007) 37:593–608.

[9] LE THI Hoai An, LE Hoai Minh, NGUYEN Van Vinh, PHAMDINHTao, Combined Feature Selection and Classification using DCA, IEEE IC on Research, Innovation and Vision for the Future (RIVF), 2008, pp:233-239.

[10] Le Thi Hoai An, Van Vinh Nguyen, Samir Ouchani: Gene Selection for Cancer Classification Using DCA. ADMA 2008: 62-72 .

[11] Le Thi Hoai An, Le Hoai Minh, Nguyen Trong Phuc, Pham Dinh Tao: Noisy Image Segmentation by a Robust Clustering Algorithm Based on DC Programming and DCA. Industrial Conf. DM 2008: 72-86.

- Thanks a lot!

- Q & A