Available online at www.sciencedirect.com

SCIENCE DIRECT®

European Journal of Operational Research 166 (2005) 756–768

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

MOP/GP models for machine learning

Hirotaka Nakayama [a,*], Ye Boon Yun [b], Takeshi Asada [c], Min Yoon [d]

[a] *Department of Applied Mathematics, Konan University, 8-9-1 Okamoto, Higashinada, Kobe 658-8501, Japan*
[b] *Kagawa University, Kagawa 761-0396, Japan*
[c] *Osaka University, Osaka 565-0871, Japan*
[d] *Yonsei University, Seoul 120-749, Republic of Korea*

Received 1 September 2002; accepted 1 March 2004
Available online 27 August 2004

Abstract

Techniques for machine learning have been extensively studied in recent years as effective tools in data mining. Although there have been several approaches to machine learning, we focus on the mathematical programming (in particular, multi-objective and goal programming; MOP/GP) approaches in this paper. Among them, Support Vector Machine (SVM) is gaining much popularity recently. In pattern classification problems with two class sets, its idea is to find a maximal margin separating hyperplane which gives the greatest separation between the classes in a high dimensional feature space. This task is performed by solving a quadratic programming problem in a traditional formulation, and can be reduced to solving a linear programming in another formulation. However, the idea of maximal margin separation is not quite new: in the 1960s the multi-surface method (MSM) was suggested by Mangasarian. In the 1980s, linear classifiers using goal programming were developed extensively.

This paper presents an overview on how effectively MOP/GP techniques can be applied to machine learning such as SVM, and discusses their problems.

Keywords: Data mining; Machine learning; Linear classifier with maximal margin; Support vector machine; Goal programming

1. Introduction

One of main purposes in data mining is to discover knowledge in data bases with very large scale. Usually, machine learning techniques are utilized for this knowledge acquisition. Typical approaches to

[*] Corresponding author. Tel.: +81 78 4352534; fax: +81 78 4525507.

E-mail addresses: nakayama@konan-u.ac.jp (H. Nakayama), yun@eng.kagawa-u.ac.jp (Y.B. Yun), asada@sa.eie.eng.osaka-u.ac.jp (T. Asada), myoon@base.yonsei.ac.kr (M. Yoon).

0377-2217/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2004.03.043

machine learning are (1) to find an explicit rule as *if-then rule* and (2) to judge newly observed data by an implicit rule which is usually represented as a nonlinear function. Well known ID3 (recently C5.0) and CART belong to the former category. On the other hand, artificial neural networks and mathematical programming approaches belong to the latter category. In this paper, we focus on the latter category.

For convenience, we consider pattern classification problems. Let X be a space of conditional attributes. For binary classification problems, the value of +1 or −1 is assigned to each data $\boldsymbol{x}_i \in X$ according to its class $\mathscr{A}$ or $\mathscr{B}$. The aim of machine learning is to predict which class newly observed data belong to on the basis of the given training data set $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $(i = 1, \ldots, l)$, where $y_i = +1$ or -1. This is performed by finding a discriminant function $f(\boldsymbol{x})$ such that $f(\boldsymbol{x}) \geqq 0$ for $\boldsymbol{x} \in \mathscr{A}$ and $f(\boldsymbol{x}) < 0$ for $\boldsymbol{x} \in \mathscr{B}$. Linear discriminant functions, in particular, can be expressed by the following linear form:

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \tag{1}$$

with the property

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \geqq 0 \quad \text{for } \boldsymbol{x} \in \mathscr{A}, \tag{2}$$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b < 0 \quad \text{for } \boldsymbol{x} \in \mathscr{B}. \tag{3}$$

Letting A denote the matrix whose row vectors are $\boldsymbol{x}_i \in \mathscr{A}$, $i = 1, \ldots, l$, then the following two notations are equivalent:

$$\text{(i)} \quad \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + b \geqq 0 \quad \text{for } \boldsymbol{x}_i \in \mathscr{A}, \ \ i = 1, \ldots, l, \tag{4}$$

$$\text{(ii)} \quad A\boldsymbol{w} + b\mathbf{1} \geqq 0 \quad \text{where } \mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}. \tag{5}$$

The latter formulation is often used in MSM, whereas the former formulation is often used in SVM and goal programming approaches, which will be described later in some detail.

For such a pattern classification problem, artificial neural networks have been widely applied. However, the back propagation method is reduced to nonlinear optimization with multiple local optima, and hence difficult to apply to large scale problems. Another drawback in the back propagation method is in the fact that it is difficult to change the structure adaptively according to the change of environment in incremental learning. Recently, Support Vector Machine (SVM, in short) is attracting interest of researchers, in particular, people who are engaged in mathematical programming, because it is reduced to quadratic programming (QP) or linear programming (LP). One of main features in SVM is that it is a linear classifier with maximal margin on the feature space. The idea of maximal margin in linear classifier has a long history in mathematical programming and goal programming. In the following in this paper, we review it in brief and discuss how effectively techniques in multi-objective programming and goal programming (MOP/GP) can be applied.

2. Multi-surface method (MSM)

Suppose that given data in a set X of n-dimensional Euclidean space belong to one of two categories $\mathscr{A}$ and $\mathscr{B}$. Let A be a matrix whose row vectors denote points of the category $\mathscr{A}$. Similarly, let B be a matrix whose row vectors denote points of the category $\mathscr{B}$. For simplicity of notation, we denote the set of points of $\mathscr{A}$ by A. The set of points of $\mathscr{B}$ is denoted by B similarly.

MSM suggested by Mangasarian (1968) finds a piecewise linear discrimination surface separating two sets A and B by solving linear programming problems iteratively. The main idea is to find two hyperplanes parallel with each other which classify as many given data as possible:

$$g(\boldsymbol{w}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} = \alpha, \tag{6}$$

$$g(\boldsymbol{w}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} = \beta. \tag{7}$$

This is performed by the following algorithm (Figs. 1–4 give a schematical implication for the algorithm):

Step 1. Solve the following linear programming problem at kth iteration (set $k = 1$ at the beginning):

$$\begin{array}{lll} (\mathrm{MSM}) & \text{Maximize} & \phi_i(A,B) = \alpha - \beta \\ & \text{subject to} & A\boldsymbol{w} \geqq \alpha\mathbf{1}, \\ & & B\boldsymbol{w} \leqq \beta\mathbf{1}, \\ & & -\mathbf{1} \leqq \boldsymbol{w} \leqq \mathbf{1}, \\ & & \boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{w} \geqq \dfrac{1}{2}\left(\dfrac{1}{2} + \boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i\right), \end{array}$$

where $\boldsymbol{p}_i$ is given by one of $\boldsymbol{p}_1^{\mathrm{T}} = (\frac{1}{\sqrt{2}}, 0, \ldots, 0)$, $\boldsymbol{p}_2^{\mathrm{T}} = (-\frac{1}{\sqrt{2}}, 0, \ldots, 0), \ldots, \boldsymbol{p}_{2n}^{\mathrm{T}} = (0, \ldots, 0, -\frac{1}{\sqrt{2}})$.

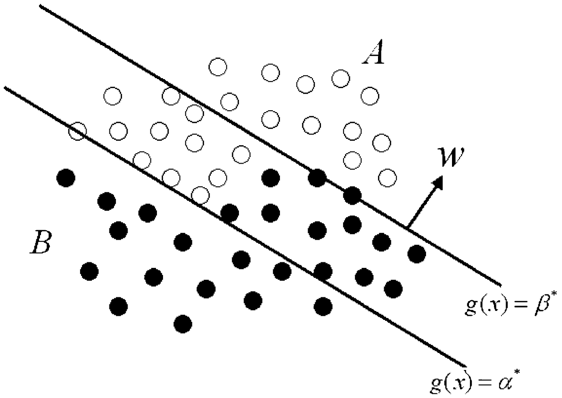

Fig. 1. MSM (Step 1).

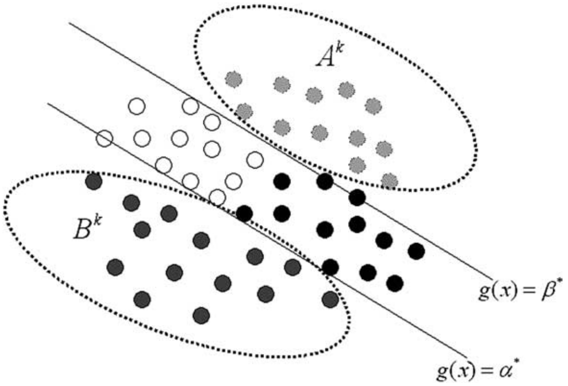

Fig. 2. MSM (Step 2).

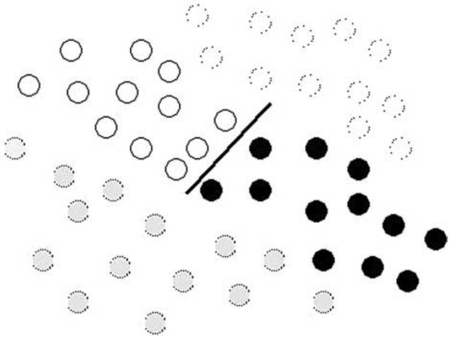

Fig. 3. MSM (Step 1–2).

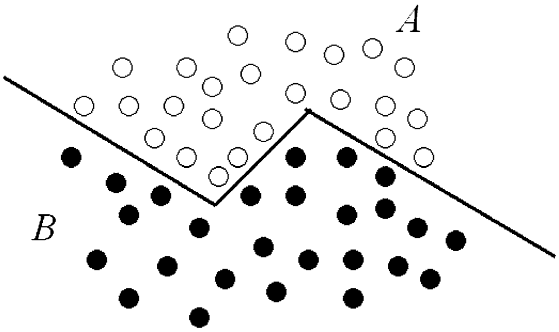

Fig. 4. MSM (Step 3).

Here the last constraint of (MSM) is introduced in order to avoid a trivial solution $\boldsymbol{w} = 0$, $\alpha = 0$, $\beta = 0$ from a linear approximation of $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \geqq \frac{1}{2}$. Namely,

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \cong \boldsymbol{p}^{\mathrm{T}}\boldsymbol{p} + 2\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{p}) \geqq \frac{1}{2}. \tag{8}$$

After solving LP problem (MSM) for each i such that $1 \leqq i \leqq 2n$, we take a hyperplane which classify correctly as many given data as possible. Let the solution be $\boldsymbol{w}^*$, α^*, β^*, and let the corresponding value of objective function be $\phi^*(A, B)$.

If $\phi^*(A, B) \geqq 0$, then we have a complete separation for A and B. We take the separating hyperplane as $g(\boldsymbol{w}^*) = (\alpha^* + \beta^*)/2$. Go to Step 3.

Otherwise (i.e., if $\phi^*(A, B) < 0$), go to Step 2.

Step 2. First, remove the points such that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}^* > \beta^*$ from the set A. Let A^k denote the set of removed points. Take the separating hyperplane as $g(\boldsymbol{w}^*) = (\beta^* + \tilde{\beta})/2$ where $\tilde{\beta} = \mathrm{Min}\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}^* \mid \boldsymbol{x} \in A^k\}$. The set A^k denotes a subregion of the category $\mathscr{A}$ in X which is decided at this stage. Rewrite $X \setminus A^k$ by X and $A \setminus A^k$ by A.

Next, remove the points such that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}^* < \alpha^*$ from the set B. Let B^k denote the set of removed points. Take the separating hyperplane as $g(\boldsymbol{w}^*) = (\alpha^* + \tilde{\alpha})/2$ where $\tilde{\alpha} = \mathrm{Max}\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}^* \mid \boldsymbol{x} \in B^k\}$. The set B^k denotes a subregion of the category $\mathscr{B}$ in X which is decided at this stage. Rewrite $X \setminus B^k$ by X and $B \setminus B^k$ by B.

Set $k = k + 1$ and go to Step 1.

Step 3. Construct a piecewise linear separating hypersurface for A and B by adopting the relevant parts of the hyperplanes obtained above.

Remark. At the final pth stage, we have the region of $\mathscr{A}$ in X as $A^1 \cup A^2 \cup \ldots \cup A^p$ and that of $\mathscr{B}$ in X as $B^1 \cup B^2 \cup \ldots \cup B^p$. Given a new point, its classification is easily made. Namely, since the new point is in either one of these subregions in X, we can classify it by checking which subregion it belongs to in the order of $1, 2, \ldots, p$.

As stated above, if $\phi^*(A,B) > 0$, then the given data set can be linearly separated. Then, note that the parallel hyperplanes $g(\boldsymbol{w}^*) = \alpha^*$ and $g(\boldsymbol{w}^*) = \beta^*$ solving LP problem (MSM) provides a maximal margin.

3. Goal programming approaches to pattern classification

MSM often provides too complex discrimination boundaries, which results in a poor ability of generalization as can be seen in Fig. 5. In 1981, Freed and Glover suggested to get just a hyperplane separating two classes with as few misclassified data as possible by using goal programming (Freed and Glover, 1981; see also Erenguc and Koehler, 1990). Let ξ_i denote the exterior deviation which is a deviation from the hyperplane of a point $\boldsymbol{x}_i$ improperly classified. Similarly, let η_i denote the interior deviation which is a deviation from the hyperplane of a point $\boldsymbol{x}_i$ properly classified. Some of main objectives in this approach are as follows:

(i) Minimize the maximum exterior deviation (decrease errors as much as possible);
(ii) Maximize the minimum interior deviation (i.e., maximize the margin);
(iii) Maximize the weighted sum of interior deviation;
(iv) Minimize the weighted sum of exterior deviation.

Although many models have been suggested, the one considering (iii) and (iv) above may be given by the following linear goal programming:

$$\begin{aligned} \text{(GP)} \quad \text{Minimize} \quad & \sum_{i=1}^{l} (h_i \xi_i - k_i \eta_i) \\ \text{subject to} \quad & y_i(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w} + b) = \eta_i - \xi_i, \\ & \xi_i, \ \eta_i \geqq 0, \ \ i = 1, \ldots, l, \end{aligned}$$

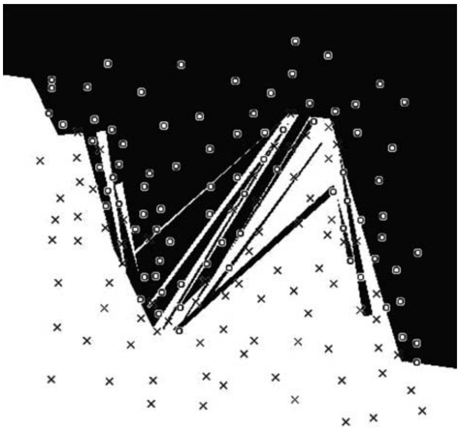

Fig. 5. Result by MSM.

where since $y_i = +1$ or -1 according to $\boldsymbol{x}_i \in \mathscr{A}$ or $\boldsymbol{x}_i \in \mathscr{B}$, two equations $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + b = \eta_i - \xi_i$ for $\boldsymbol{x}_i \in \mathscr{A}$ and $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + b = -\eta_i + \xi_i$ for $\boldsymbol{x}_i \in \mathscr{B}$ can be reduced to the following one equation:

$$y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + b) = \eta_i - \xi_i. \tag{9}$$

Here h_i and k_i are positive constants. In order for ξ_i and η_i to have the meaning of the exterior deviation and the interior deviation respectively, the condition $\xi_i\eta_i = 0$ for every $i = 1,\ldots,l$ must hold.

Lemma 1. *If $h_i > k_i$ for $i = 1,\ldots,l$, then we have $\xi_i\eta_i = 0$ for every $i = 1,\ldots,l$ at the solution to (GP).*

Proof. Easy due to Lemma 7.3.1 of Sawaragi et al. (1994). □

It should be noted that the above formulation may yield some unacceptable solutions such as $\boldsymbol{w} = 0$ and unbounded solution.

Example. Let $\boldsymbol{x}_1 = (-1,1)$, $\boldsymbol{x}_2 = (0,2) \in \mathscr{A}$ and $\boldsymbol{x}_3 = (1,-1)$, $\boldsymbol{x}_4 = (0,-2) \in \mathscr{B}$. Constraint functions of (GP) are given by

$$\boldsymbol{x}_1 : w_1(-1) + w_2(1) + b = \eta_1 - \xi_1, \tag{10}$$

$$\boldsymbol{x}_2 : w_1(0) + w_2(2) + b = \eta_2 - \xi_2, \tag{11}$$

$$\boldsymbol{x}_3 : w_1(1) + w_2(-1) + b = -\eta_3 + \xi_3, \tag{12}$$

$$\boldsymbol{x}_4 : w_1(0) + w_2(-2) + b = -\eta_4 + \xi_4. \tag{13}$$

Here it is clear that $\xi = 0$ at the optimal solution. Note that the feasible (w_1,w_2) in (w_1,w_2)-space moves to the north–west by increasing η_i (see Fig. 6). Maximizing $\sum\eta_i$ yields unbounded optimal solution unless any further constraint for $\boldsymbol{w}$ are added. In the goal programming approach to linear classifiers, therefore, some appropriate normality condition must be imposed on $\boldsymbol{w}$ in order to provide a bounded nontrivial optimal solution. One of such normality conditions is $\|\boldsymbol{w}\| = 1$.

If the classification problem is linearly separable, then using the normalization $\|\boldsymbol{w}\| = 1$, the separating hyperplane H: $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$ with maximal margin can be given by solving the following problem (Cavalier et al., 1989):

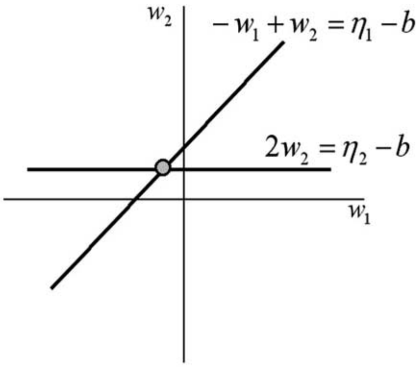

Fig. 6. A cause of unacceptable solution.

$$\begin{array}{lll} (\mathrm{GP}_1) & \text{Maximize} & \eta \\ & \text{subject to} & y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b) \geqq \eta, \quad i=1,\ldots,l, \\ & & \|\boldsymbol{w}\| = 1. \end{array}$$

However, this normality condition makes the problem to be of nonlinear optimization. Instead of maximizing the minimum interior deviation in (GP_1), we can use the following equivalent formulation with the normalization $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=\pm 1$ at points with the minimum interior deviation (Marcotte and Savard, 1992):

$$\begin{array}{lll} (\mathrm{GP}_1') & \text{Minimize} & \|\boldsymbol{w}\| \\ & \text{subject to} & y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b) \geqq \eta, \quad i=1,\ldots,l, \\ & & \eta = 1. \end{array}$$

This formulation is the same as the one used in Support Vector Machine (SVM) which will be stated later.

4. Revision of MSM by MOP/GP

One of drawbacks in MSM is the fact that it yields sometimes too complex discrimination boundaries which cause poor generalization ability. In Nakayama and Kagaku, 1998, several modifications of MSM are suggested. One of them introduces interior deviations as well as exterior deviations in MSM. This is formulated as a multi-objective programming. If only exterior deviations are considered, this is reduced to a goal programming, which is the same as the one suggested by Bennett and Mangasarian (1992) called RLPD (robust linear programming discrimination). Applying these MOP/GP approaches to MSM, we can obtain smoother discrimination boundary than the original MSM.

Furthermore, Nakayama and Kagaku (1998) applied a fuzzy programming technique to MSM, because it is more natural to regard the constraints $A\boldsymbol{w}+b\mathbf{1} \geqq \mathbf{0}$ and $B\boldsymbol{w}+b\mathbf{1} \leqq \mathbf{0}$ as those which are to be satisfied approximately. For example, the constraints $A\boldsymbol{w}+b\mathbf{1} \geqq \mathbf{0}$ may be fuzzified into

$$A\boldsymbol{w}+b\mathbf{1} \succeq 0, \tag{14}$$

where $\succeq$ implies "almost $\geqq$". The degree of satisfaction of the inequality $A\boldsymbol{w}+b\mathbf{1} \geqq 0$ can be represented by the membership function $M_A(\boldsymbol{x})$ denoting the degree of membership of $\boldsymbol{x} \in A$:

$$M_A(\boldsymbol{x}) = \begin{cases} 1, & \frac{\mathbf{x}^{\mathrm{T}}\mathbf{w}+b}{e} \geqq 1, \\ \frac{1}{2}\frac{\mathbf{x}^{\mathrm{T}}\mathbf{w}+b}{e}+\frac{1}{2}, & -1 < \frac{\mathbf{x}^{\mathrm{T}}\mathbf{w}+b}{e} < 1, \\ 0, & \frac{\mathbf{x}^{\mathrm{T}}\mathbf{w}+b}{e} \leqq -1. \end{cases} \tag{15}$$

This approach yields gray zones for discrimination boundaries, in which the data are not decided clearly as of $\mathscr{A}$ or $\mathscr{B}$. However, this is rather natural in practical situations, because we usually require further investigation on those data as in cases of medical diagnosis. The parameter e represents the width of the "gray zone" of the separating hyperplane. The fuzziness of the boundary between the set A and B varies according to the value of these parameters. They should be decided by users on the basis of their experiences.

5. Support vector machine

Support vector machine (SVM) is developed by Vapnik (1995) (see also Cristianini and Shawe-Taylor, 2000), and its main features are

(1) SVM is based on linear classifiers with maximal margin on the feature space,
(2) SVM uses kernel representation preserving inner products on the feature space,
(3) SVM provides an evaluation of the generalization ability using VC dimension.

In cases where training data set X is not linearly separable, we map the original data set X to a feature space Z by some nonlinear map ϕ. Increasing the dimension of the feature space, it is expected that the mapped data set becomes linearly separable. We try to find linear classifiers with maximal margin in the feature space. Letting $\boldsymbol{z}_i = \phi(\boldsymbol{x}_i)$, the separating hyperplane with maximal margin can be given by solving the following problem with the normalization $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z} + b = \pm 1$ at points with the minimum interior deviation:

$$(\mathrm{SVM}_{\mathrm{hard}}) \quad \text{Minimize} \quad \|\boldsymbol{w}\|$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_i + b) \geqq 1, \quad i = 1, \ldots, l.$$

Several kinds of norm are possible. When $\|\boldsymbol{w}\|_2$ is used, the problem is reduced to quadratic programming, while the problem with $\|\boldsymbol{w}\|_1$ or $\|\boldsymbol{w}\|_\infty$ is reduced to linear programming (see, e.g., Mangasarian, 2000).

For the above example, we have the following condition in the SVM formulation:

$$z_1 : w_1(-1) + w_2(1) + b \geqq 1, \tag{16}$$

$$z_2 : w_1(0) + w_2(2) + b \geqq 1, \tag{17}$$

$$z_3 : w_1(1) + w_2(-1) + b \leqq -1, \tag{18}$$

$$z_4 : w_1(0) + w_2(-2) + b \leqq -1. \tag{19}$$

Since it is clear that the optimal hyperplane has $b = 0$, the constraint functions for $z_3$and z_4 are identical to those for z_1 and z_2. The feasible region in (w_1, w_2)-plane is given by $w_2 \geqq w_1 + 1$ and $w_2 \geqq 1/2$. Minimizing the objective function of SVM yields the optimal solution $(w_1, w_2) = (-1/2, 1/2)$ for the QP formulation (see Fig. 7). Similarly, we have a solution among the line segment $\{w_2 \geqq w_1 + 1\} \cap \{-1/2 \leqq w_1 \leqq 0\}$ depending on the initial solution for the LP formulation.

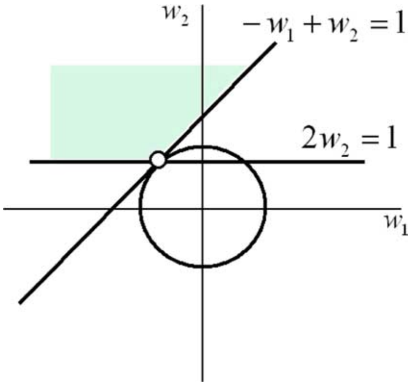

Fig. 7. SVM gets rid of unacceptable solution.

Dual problem of ($\mathrm{SVM_{hard}}$) with $\|\boldsymbol{w}\|_2$ is

$$(\mathrm{SVM'_{hard}})\quad \text{Maximize}\quad \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

$$\text{subject to}\quad \sum_{i=1}^{l}\alpha_i y_i=0,$$

$$\alpha_i \geqq 0,\quad i=1,\dots,l.$$

Using the kernel function $K(\boldsymbol{x},\boldsymbol{x}')=\phi(\boldsymbol{x})^{\mathrm{T}}\ \phi(\boldsymbol{x}')$, the problem ($\mathrm{SVM'_{hard}}$) can be reformulated as follows:

$$(\mathrm{SVM''_{hard}})\quad \text{Maximize}\quad \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$\text{subject to}\quad \sum_{i=1}^{l}\alpha_i y_i=0,$$

$$\alpha_i \geqq 0,\quad i=1,\dots,l.$$

Several kinds of kernel functions have been suggested: among them, q-polynomial

$$K(\boldsymbol{x},\boldsymbol{x}')=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'+1)^q \tag{20}$$

and Gaussian

$$K(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-\frac{\|\ \boldsymbol{x}-\boldsymbol{x}'\|^2}{r^2}\right) \tag{21}$$

are most popularly used. In applying the Gaussian kernel, it is important to decide the parameter r. The author and his coresearchers have observed through their numerical experiments that the value of r may be effectively determined by the simple estimate modifying the formula given by Haykin (1994) slightly,

$$r=\frac{d_{\max}}{\sqrt[n]{nl}}, \tag{22}$$

where $d_{\max}$ is the maximal distance among the data; n is the dimension of data; l is the number of data.

Unlike MSM and original GP-approaches, SVM can provide smooth nonlinear discrimination boundaries in the original data space which result in better generalization ability. However, it can be expected that many devices in MSM and MOP/GP approaches to linear classifiers can be applied to SVM.

5.1. Hard margin and soft margin

Separating two sets A and B completely is called the hard margin method, which tends to make overlearning. This implies the hard margin method is easily affected by noise. In order to overcome this difficulty, the soft margin method is introduced. The soft margin method allows some slight error which is represented by slack variables (exterior deviation) $\xi_i\ (i=1,\dots,l)$. Using the trade-off parameter C between minimizing $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ and minimizing $\sum_{i=1}^{l}\xi_i$, we have the following formulation for the soft margin method:

$$(\mathrm{SVM_{soft}})\quad \text{Minimize}\quad \|\boldsymbol{w}\|+C\sum_{i=1}^{l}\xi_i$$

$$\text{subject to}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_i+b)\geqq 1-\xi_i,$$

$$\xi_i\geqq 0,\quad i=1,\dots,l.$$

Using a kernel function in the dual problem yields

$$
\begin{aligned}
(\mathrm{SVM}'_{\mathrm{soft}})\quad & \text{Maximize} && \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\\
& \text{subject to} && \sum_{i=1}^{l}\alpha_i y_i = 0,\\
& && C \geqq \alpha_i \geqq 0, \quad i = 1,\ldots,l.
\end{aligned}
$$

It can be seen that the idea of soft margin method is the same as the goal programming approach to linear classifiers. Not only exterior deviations but also interior deviations can be considered in SVM. Such MOP/GP approaches to SVM are discussed by the author and his coresearchers (Nakayama and Asada, 2001; Asada and Nakayama, 2003; Yoon et al., 2003). When applying GP approaches, it was pointed out in Section 3 that we need some normality condition in order to avoid unacceptable solutions.

Glover (1990) suggested the following necessary and sufficient condition for avoiding unacceptable solutions:

$$
\left(-l_A\sum_{i\in I_B}\boldsymbol{x}_i + l_B\sum_{i\in I_A}\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{w} = 1, \tag{23}
$$

where l_A and l_B denote the number of data for the category $\mathscr{A}$ and $\mathscr{B}$, respectively.

Geometrically, the normalization (23) means that the distance between two hyperplanes passing through centers of data for l_A and l_B is scaled by $l_A\ l_B$. Taking into account that $\eta_i/\|\boldsymbol{w}\|$ represents the margin of correctly classified data $\boldsymbol{x}_i$ from the hyperplane $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$, larger value of η_i and smaller value of $\|\boldsymbol{w}\|$ are more desirable in order to maximize the margin. On the other hand, since $\xi_i/\|\boldsymbol{w}\|$ stands for the margin of misclassified data, the value of ξ_i should be minimized. The methods considering all of $\xi_i/\|\boldsymbol{w}\|$ and $\eta_i/\|\boldsymbol{w}\|$ are referred to as the total margin methods. Now, we have the following formulation for getting a linear classifier with maximal total margin:

$$
\begin{aligned}
(\mathrm{SVM}_{\mathrm{total}})\quad & \text{Minimize} && \|\boldsymbol{w}\| + \sum_{i=1}^{l}(h_i\xi_i - k_i\eta_i)\\
& \text{subject to} && y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_i + b) = 1 + \eta_i - \xi_i,\\
& && \xi_i,\ \ \eta_i \geqq 0,\ \ i = 1,\ldots,l.
\end{aligned}
$$

Here the condition $h_i > k_i$, $i = 1,\ldots,l$, is imposed in order to assure $\xi_i\eta_i = 0$, $i = 1,\ldots,l$.

Figs. 8–10 show some graphical implication of comparison among GP, $\mathrm{SVM}_{\mathrm{hard}}$, $\mathrm{SVM}_{\mathrm{soft}}$ and $\mathrm{SVM}_{\mathrm{total}}$. An error bound for generalization of $\mathrm{SVM}_{\mathrm{total}}$ is given in Yoon et al. (2003).

6. Problems in machine learning

SVM is an elegant method for machine learning using linear classifiers with maximal margin and kernel representation, which has been widely applied to practical problems. However, there remain several problems to be resolved not only in SVM but also in other methods as machine learning methodologies:

(1) unbalance in data sets,
(2) outlier or not?
(3) additional learning.

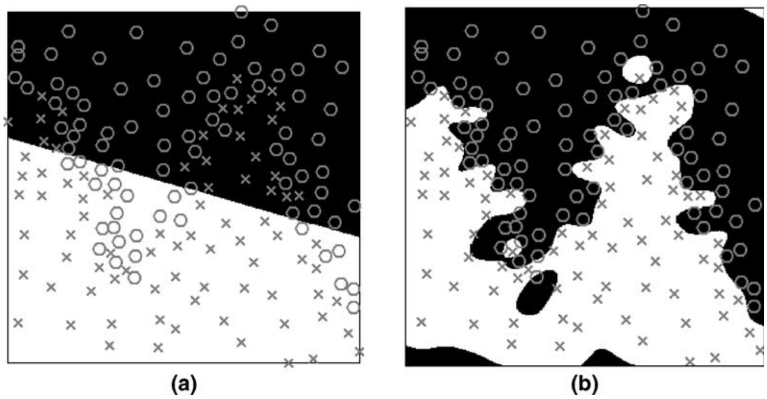

Fig. 8. Results by goal programming and hard margin SVM: (a) GP and (b) $\mathrm{SVM}_{\mathrm{hard}}$.

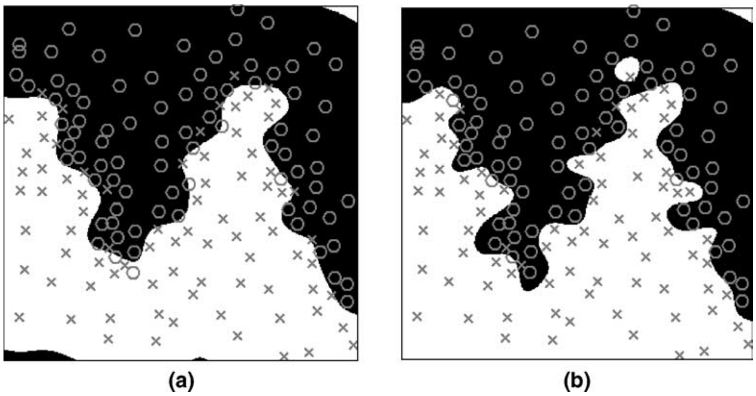

Fig. 9. Results by soft margin SVM: (a) $C = 1$ and (b) $C = 10$.

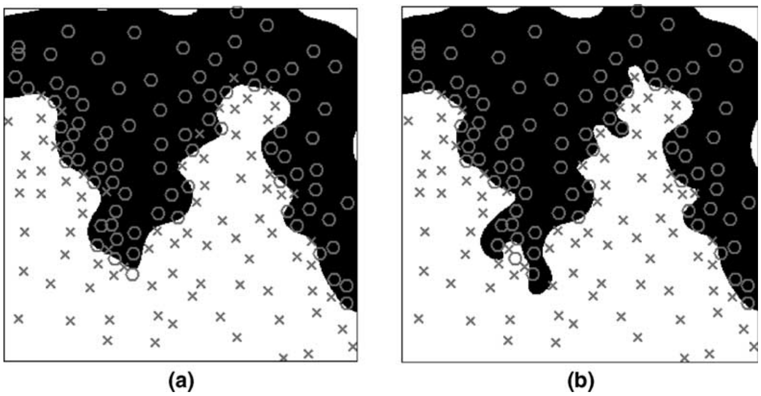

Fig. 10. Results by total margin SVM: (a) $h_i = 10$, $k_i = 5$, $\forall i$ and (b) $h_i = 100$, $k_i = 5$, $\forall i$.

6.1. Unbalance of data

The author and his coresearchers are tackling with the problem for predicting land-slide disasters on the basis of land shape, type of soil/rock, amount of rain fall, etc. Although the data base is very huge, the data for disaster occurrence is very few (usually less than 1.0%). This causes a difficulty in predicting a newly observed sample to be disastrous. It is disastrous data that are needed to be judged correctly. Similar situations happen in forecasting bankruptcy. Asada and Nakayama (2003) tries to enlarge the region of the category with fewer data by minimizing the sum of interior deviations. Other methods such as boosting, which try to make learning for misclassified data more strongly, should be examined in SVM in the future.

6.2. Outlier or not?

Sometimes isolated data are judged to be outliers. Of course, whether or not the data is outlier depends on the amount of noise. In addition, we have to note another important situation. That is the change of environment. Many problems are time dependent. Even though a sample is isolated at a moment, many samples of the same category as that of the isolated one may appear in the neighborhood of the isolated sample as the time passes. In this event, the isolated sample is not outlier, but an important sample for future.

One problem is that we cannot know at the moment whether the isolated sample is an outlier or not. Therefore, it is better to get a separating hyperplane taking into account both cases in which the isolated sample is an outlier and an important one. This can be realized by introducing the interior deviation η in SVM (Asada and Nakayama, 2003), or the constraint $\xi_i \leqq \xi_{\max}$ (Yoon et al., 2003).

6.3. Additional learning

The environment of decision making changes over time. Therefore, we have to revise knowledge obtained from data mining according to the change of environment. To this end, additional learning becomes an important task in machine learning. Additional learning has been studied extensively in machine learning and artificial neural networks, e.g. Platt (1991), Yamauchi et al. (1999), Nakayama and Yoshii (2000). It should be noted that if we make only additional learning, the obtained rule becomes more and more complex which results in poor generalization ability. In order to make rules simple, we have to remove unnecessary data. This is forgetting. A simple way for forgetting a sample is to decrease the influence of the sample as the time passes. This is called passive forgetting. However, some old data are important, while some new data are less important. It is needed, therefore, to find unnecessary data more actively. The author and his coresearchers have reported a way for active forgetting and its effectiveness along an example of stock portfolio problems (Nakayama and Yoshii, 2000; Nakayama and Asada, 2001; Nakayama and Hattori, 2002).

7. Concluding remarks

A brief survey of mathematical programming (in particular, MOP/GP) models for machine learning was presented in this paper. In recent years, SVM has been gaining much popularity. It was shown that MOP/GP techniques can be effectively applied to machine learning such as SVM. They are expected to be effective in particular for problems with unbalance among data set and for problems under the change of environment. Further researches in applications shall be needed for practical implications in this field in the future.

References

Asada, T., Nakayama, H., 2003. SVM using multiobjective linear programming and goal programming. In: Tanino, T., Tanaka, T., Inuiguchi, M. (Eds.), Multi-objective Programming and Goal Programming. Springer, Berlin, pp. 93–98.

Bennett, K.P., Mangasarian, O.L., 1992. Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software 1, 23–34.

Cavalier, T.M., Ignizio, J.P., Soyster, A.L., 1989. Discriminant analysis via mathematical programming: Certain problems and their causes. Computers and Operations Research 16, 353–362.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge.

Erenguc, S.S., Koehler, G.J., 1990. Survey of mathematical programming models and experimental results for linear discriminant analysis. Managerial and Decision Economics 11, 215–225.

Freed, N., Glover, F., 1981. Simple but powerful goal programming models for discriminant problems. European Journal of Operational Research 7, 44–60.

Glover, F., 1990. Improved linear programming models for discriminant analysis. Decision Sciences 21, 771–785.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Macmillan, London.

Mangasarian, O.L., 1968. Multisurface method of pattern separation. IEEE Transactions on Information Theory IT-14, 801–807.

Mangasarian, O.L., 2000. Generalized support vector machines. In: Smola, A., Bartlett, P., Shölkopf, B., Schuurmans, D. (Eds.), Advances in Large Margin Classifiers. MIT Press, Cambridge, MA, pp. 135–146.

Marcotte, P., Savard, G., 1992. Novel approaches to the discrimination problem. ZOR—Methods and Models of Operations Research 36, 517–545.

Nakayama, H., Asada, T., 2001. Support vector machines formulated as multiobjective linear programming. Proceedings of ICOTA 3, 1171–1178.

Nakayama, H., Hattori, A., 2002. Additional learning and forgetting by support vector machine and RBF networks. Proceedings of ICONIP, in CD-ROM.

Nakayama, H., Kagaku, N., 1998. Pattern classification by linear goal programming and its extensions. Journal of Global Optimization 12, 111–126.

Nakayama, H., Yoshii, K., 2000. Active forgetting in machine learning and its application to financial problems. Proceedings of International Joint Conference on Neural Networks, in CD-ROM.

Platt, J., 1991. A resource allocating network for function interpolation. Neural Computation 3, 213–225.

Sawaragi, Y., Nakayama, H., Tanino, T., 1994. Theory of Multiobjective Optimization. Academic Press, New York.

Vapnik, V.P., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.

Yamauchi, K., Yamaguchi, N., Ishii, N., 1999. Incremental learning methods with retrieving of interfered patterns. IEEE Transactions on Neural Networks 10–6, 1351–1365.

Yoon, M., Nakayama, H., Yun, Y.B., 2003. A Soft Margin Algorithm controlling Tolerance Directly. In: Tanino, T., Tanaka, T., Inuiguchi, M. (Eds.), Multi-objective Programming and Goal Programming, pp. 281–288.

Yoon, M., Yun, Y.B., Nakayama, H., 2003. A role of total margin in support vector machines. Proceedings of IJCNN'03, 2049–2053.