The Korean Communications in Statistics Vol. 14 No. 1, 2007, pp. 133–140

A Note on Fuzzy Support Vector Classification*

Sungho Lee[1] and Dug Hun Hong[2]

Abstract

The support vector machine has been well developed as a powerful tool for solving classification problems. In many real world applications, each training point has a different effect on constructing classification rule. Lin and Wang (2002) proposed fuzzy support vector machines for this kind of classification problems, which assign fuzzy memberships to the input data and reformulate the support vector classification. In this paper another intuitive approach is proposed by using the fuzzy $\alpha-$cut set. It will show us the trend of classification functions as α changes.

Keywords: SVM; SVC; fuzzy membership; $\alpha-$cut set.

1. Introduction

The support vector machine (SVM) is a tool for solving multidimensional function estimation problems. It was developed in Russia in the sixties by Vapnik and co-workers (Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1964). It was initially designed to solve pattern recognition problems, where one selects some (small) subset of the training data, called the support vectors, to find a decision rule with good generalization ability. Later the support vector machine was extended to regression and real-valued function estimation. The support vector machine is a very powerful method in a wide variety of applications and has been introduced as a powerful tool for solving classification problems (Vapnik, 1995, 1998; Gun, 1998; Schölkopf and Smola, 2002).

The support vector classification (SVC) algorithm maps the input data in $\mathbb{R}^m$ into a high dimensional feature space F with a dot product, *i.e.*, $\phi : \mathbb{R}^m \to F$ and finds the optimal hyperplane to maximize the margin between two classes

* This work was supported (in part) by the Daegu University research Grant, 2006.

[1] Professor, Department of Statistics, Daegu University, Kyungbug 712-714, Korea.
Correspondence : shlee1@daegu.ac.kr

[2] Professor, Department of mathematics, Myongji University, Kyunggi 449-728, Korea.

in F. The optimization problem can be transformed into its corresponding dual problem by using the Lagrangian multipliers and is reduced to a quadratic programming problem formulated in terms of dot products in F. Instead of evaluating the map, $\phi : \mathbb{R}^m \to F$ explicitly, a kernel function $k(x, y)$ is used to compute a dot product in feature space F, *i.e.*, $k(x, y) = (\phi(x), \phi(y))$ and the solution to the optimization problem is given by a function of support vectors.

In many real world classification problems, some training points are more important than others and some are less meaningful. For example, some points contaminated by errors or noises are less meaningful than others. Thus it is not reasonable that those training points have the same weights as others. Lin and Wang (2002) proposed fuzzy support vector machines for this kind of classification problems. The proposed method assigns fuzzy memberships to the input data and reformulates the support vector classification algorithm such that different input data can make different contributions to the learning of decision function. In this paper another intuitive approach is suggested by using the concept of $\alpha-$cut of fuzzy membership to solve this kind of classification problems. It will use $\alpha-$cut sets of the training points to find classification functions and hence show us the trend of classification functions as α changes.

2. SVC and Fuzzy SVC

In this section we briefly review support vector classification (Cortes and Vapnik, 1995; Vapnik, 1995, 1998; Gun, 1998; Schölkopf and Smola, 2002) and fuzzy support vector classification proposed by Lin and Wang (2002).

Let $\{(x_i, y_i) | x_i \in \mathbb{R}^m, y_i \in \{-1, 1\}, i = 1, 2, \dots, n\}$ be a training set. The main idea of support vector classification is to find a hyperplane, $(w, x) + b = 0$, to separate the two classes so that the margin (the distance between the hyperplane and the nearest point) is maximized. The optimization problem can be constructed as follows:

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C \sum_{i=1}^{n} \xi_i$$

with constraints

$$\begin{aligned} y_i((w, x_i) + b) &\geq 1 - \xi_i, \ i = 1, \dots, n, \\ \xi_i &\geq 0, \ i = 1, \dots, n, \end{aligned} \tag{2.1}$$

where $C(\geq 0)$ is a constant and $\xi_1, \dots, \xi_n$ are a measure of the misclassification errors.

The problem can be transformed into its dual problem by using the Lagrangian multipliers as follows:

$$\text{Maximize} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i, x_j)$$

with constraints

$$\begin{aligned} 0 \le \alpha_i &\le C, \ i = 1, \dots, n, \\ \sum_{i=1}^{n} \alpha_i y_i &= 0. \end{aligned}$$

In the case where a linear boundary is not appropriate the SVM maps the input data in $\mathbb{R}^m$ into a high dimensional feature space F with a dot product, *i.e.*, $\phi : \mathbb{R}^m \to F$ and finds the optimal hyperplane, $(w, \phi(x)) + b = 0$, to maximize the margin between two classes in F. The optimization problem can be constructed as follows:

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C\sum_{i=1}^{n} \xi_i$$

with constraints

$$\begin{aligned} y_i((w, \phi(x_i)) + b) &\ge 1 - \xi_i, \ i = 1, \dots, n, \\ \xi_i &\ge 0, \ i = 1, \dots, n. \end{aligned} \tag{2.2}$$

The problem can be transformed into its dual problem by using the Lagrangian multipliers as follows:

$$\begin{aligned} \text{Maximize} \quad W(\alpha) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\phi(x_i), \phi(x_j)) \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \end{aligned}$$

with constraints

$$\begin{aligned} 0 \le \alpha_i &\le C, \ i = 1, \dots, n, \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

where $k(x_i, x_j)$ is a kernel function to compute a dot product in feature space (see Vapnik, 1995).
The solution to this dual problem gives the classification function,

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right).$$

The remained question is which functions $k(x, y)$ correspond to a dot product in some feature space F. Mercer theorem (1909) indicates that any continuous symmetric function $k(x, y)$ may be used as an admissible support vector kernel (Mercer kernel) if it satisfies Mercer's condition

$$\iint k(x, y)g(x)g(y)dxdy \geq 0 \quad \text{for all} \quad g \in L_2(\mathbb{R}^m).$$

The fuzzy SVC method by Lin and Wang (2002) assigns a fuzzy membership s_i to each input point and reformulate the SVC algorithm. Let

$$\{(x_i, y_i, s_i) | x_i \in \mathbb{R}^m, y_i \in \{-1, 1\}, 0 < \sigma \leq s_i \leq 1, i = 1, 2, \ldots, n\}$$

be a training set such that the fuzzy membership s_i is the attitude of the corresponding point x_i toward one class and is larger than sufficiently small membership $\sigma > 0$. Then the optimization problem can be constructed as follows:

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C\sum_{i=1}^{n} s_i \xi_i$$

with constraints

$$\begin{aligned} y_i((w, x_i) + b) &\geq 1 - \xi_i, \ i = 1, \ldots, n, \\ \xi_i &\geq 0, \ i = 1, \ldots, n, \end{aligned} \tag{2.3}$$

where $C(\geq 0)$ is a constant and the term $s_i\xi_i$ is regarded as a measure of error with different weight in this model.

The problem can be also transformed into its dual problem by using the Lagrangian multipliers as follows:

$$\text{Maximize} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i, x_j)$$

with constraints

$$\begin{aligned} 0 \leq \alpha_i &\leq s_i C, \; i = 1, \ldots, n, \\ \sum_{i=1}^{n} \alpha_i y_i &= 0. \end{aligned}$$

In the case where a linear boundary is not appropriate the optimization problem can be constructed as before:

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C \sum_{i=1}^{n} s_i \xi_i$$

with constraints

$$\begin{aligned} y_i((w, \phi(x_i)) + b) &\geq 1 - \xi_i, \; i = 1, \ldots, n, \\ \xi_i &\geq 0, \; i = 1, \ldots, n. \end{aligned} \tag{2.4}$$

The problem can be also transformed into its dual problem by using the Lagrangian multipliers as follows:

$$\begin{aligned} \text{Maximize} \quad W(\alpha) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\phi(x_i), \phi(x_j)) \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \end{aligned}$$

with constraints

$$\begin{aligned} 0 \leq \alpha_i &\leq s_i C, \; i = 1, \ldots, n, \\ \sum_{i=1}^{n} \alpha_i y_i &= 0, \end{aligned}$$

3. Fuzzy SVC with $\alpha-$cut sets

When each training point has a different effect on constructing a classification function, the fuzzy support vector classification method was proposed by Lin and Wang (2002) to find the rule as briefly described in section 2. In this section another approach is proposed by using the fuzzy $\alpha-$cut sets, which will show us the trend for the classification function as α changes. Let

$$A = \{(x_i, y_i, s_i) | x_i \in \mathbb{R}^m, y_i \in \{-1, 1\}, 0 < s_i \leq 1, i = 1, \ldots, n\}$$

be a training set. The fuzzy membership s_i is the attitude of the corresponding point x_i toward one class as before. We define $[A]^\alpha$, $\alpha-$cut set of A, as

$$[A]^\alpha = \{(x_i, y_i, s_i) | x_i \in \mathbb{R}^m, y_i \in \{-1, 1\}, 0 < \alpha \leq s_i \leq 1, i = 1, \ldots, n\}.$$

For each $\alpha-$cut set $[A]^\alpha$ find a hyperplane, $(w, x) + b = 0$, to maximize the margin between two classes. Then the model (2.3) in section 2 can be applied to the optimization problem for each $\alpha-$cut set. That is, for $\alpha-$cut set $[A]^\alpha$

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C\sum_{i=1}^{n} s_i\xi_i$$

with constraints

$$\begin{aligned} y_i((w, x_i) + b) &\geq 1 - \xi_i, \ i = 1, \ldots, n, \\ \xi_i &\geq 0, \ i = 1, \ldots, n. \end{aligned}$$

In the case where a linear boundary is not appropriate, model (2.4) in section 2 can be used:

$$\text{Minimize} \quad \Phi(w, \xi) = \frac{1}{2}(w, w) + C\sum_{i=1}^{n} s_i\xi_i$$

with constraints

$$\begin{aligned} y_i((w, \phi(x_i)) + b) &\geq 1 - \xi_i, \ i = 1, \ldots, n, \\ \xi_i &\geq 0, \ i = 1, \ldots, n. \end{aligned}$$

A procedure for looking into the trend of classification functions is: (i) first choose a lower bound $\sigma(\leq \alpha)$, (ii) find a membership function suitable for the training data set, (iii) construct classification functions from $\alpha = 0.9$ to the smallest level σ with decreasing by 0.1 and investigate the trend for each classification function,

4. Numerical Study

In this section numerical illustrations for the results of FSVC with $\alpha-$cut sets are provided. The training data set D, for the comparison of results,

$$\begin{aligned} D &= \{(x_{1i}, x_{2i}, y_i) | (x_{1i}, x_{2i}) \in R^2, y_i \in \{-1, 1\}, i = 1, \ldots, 40\} \\ &= \{(-3.5, -2.5, -1), (-4.5, -2.0, -1), (-5.5, -0.2, 1), \ldots, (3.5, -0.5, -1)\} \end{aligned}$$

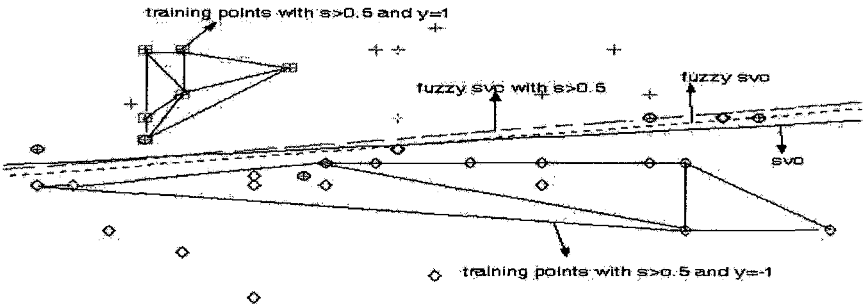
training points with s>0.5 and y=1
fuzzy svc with s>0.5
fuzzy svc
svc
training points with s>0.5 and y=-1

Figure 4.1: Classification function for the first membership function

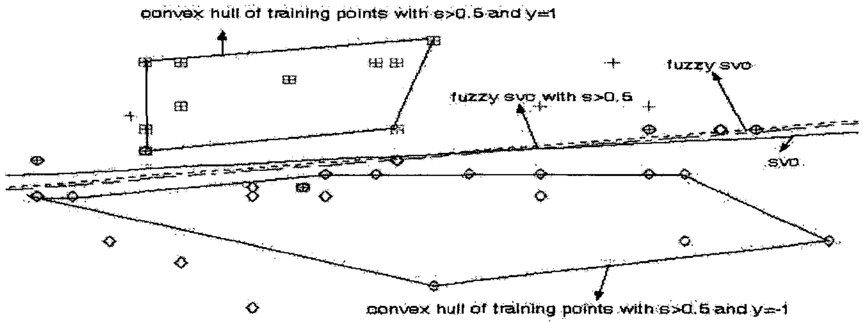
convex hull of training points with s>0.5 and y=1
fuzzy svc
fuzzy svc with s>0.5
svc
convex hull of training points with s>0.5 and y=-1

Figure 4.2: Classification functions for the second membership function

is similarly constructed based on Fig. 1 in Lin and Wang (2002). Membership functions, which is used in Lin and Wang (2002), are considered. The membership functions assign more weighting for recent points than past points:

$$s_i = f(t_i) = (1-\sigma)\left(\frac{t_i - t_1}{t_n - t_1}\right)^2 + \sigma,$$

$$s_i = f(t_i) = \frac{(1-\sigma)}{t_n - t_1}t_i + \frac{t_n\sigma - t_1}{t_n - t_1},$$

where $t_1 = i, i = 1, \ldots, 40$ and $\sigma = 0.1$.

For the support vector classifications, Gunn's program is used (see Gunn, 1998) and $C = 100$, for all the classification functions in figure 4.1 and 4.2, is chosen so that svc line in figure 4.1 may be a possible optimal separating line. Of course, $C = 100$ can't be an optimal value to satisfy all the classification lines in any sense. In experiments $C = 100, 90, \ldots, 10, 5, 1$ was applied to the model. The width of the margin, $2/||w||$, was increased from 0.627 to 0.905 and slopes of svc lines were also changed a little bit. In figure 4.1 and 3.2 the point x_i with $y_i = 1$ is placed as plus sign marker and the point x_i with $y_i = -1$ as diamond marker. Figure 4.1 shows classification functions for the first membership function using SVC, FSVC, and FSVC with $\alpha-$cut set ($\alpha = 0.5$) and figure 4.2 shows classification functions for the second membership function using SVC, FSVC, and FSVC with $\alpha-$cut set ($\alpha = 0.5$). For this training data set three classification functions are very similar but shows a trend among them.

5. Concluding Remarks

In this paper, for classification problems in which each point has different effect on classification rule, fuzzy SVC with $\alpha-$cut sets is proposed to look into the trend for classification functions as α changes. Classification functions by membership levels are considered to be able to give us informations about trend, in what directions classifiers move as the figures in the above experiment indicate. Numerical examples don't show interesting facts, but it still give us valuable informations for the trend of classification functions.

References

Cortes, C. and Vapnik, V. (1995). Support vector network. *Machine Learning*, **20**, 273–297.

Gunn, S. (1998). Support vector machines for classification and regression. *Technical report, Image Speech and Intelligent Systems Research Group*, University of Southampton.

Lin, C. F. and Wang, S. D. (2002). Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, **13**, 464–471.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transaction of the Royal Society of London*, Ser. A, **209**, 415–446.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

Vapnik, V. N. and Chervonenkis, A. J. (1964). A note on a class of perceptrons. *Automation and Remote Control*, **25**, 112–120.

Vapnik, V. and Lerner. L. (1963). Pattern Recognition using generalized portrait method. *Automation and Remote Control*, **24**, 774–780.

[Received October 2006, Accepted December 2006]