A-17

On the Influence of Outliers in the Soft Margin SVM Framework

戴 陽* 田中 純一† 渡辺 治‡
Yang Dai Jun'ichi Tanaka Osamu Watanabe

1 Introduction and Our Main Result

The *Support Vector Machine* (SVM in short) is a modern mechanism for two-class classification, regression, and clustering problems. Since the present form of SVM was proposed [5], SVMs have been used in various application areas, and their classification power has been investigated in depth from both experimental and theoretical points of view. (See, e.g., [6].) An important feature is that their way of working, by identifying the so-called support vectors among the data, offers important contributions to a number of problems related to Data Mining.

Here we consider using SVM for the two-class classification problem. A basic formulation for the problem is presented as follows. Suppose that we are given a set of m examples $\boldsymbol{x}_i$, $1 \leq i \leq m$, in an n-dimensional space, say $\mathbb{R}^n$. Each example $\boldsymbol{x}_i$ is labeled by $y_i \in \{1, -1\}$ denoting the classification of the example. Then the *SVM training problem* is essentially to solve the following optimization problem (P1). (Here we follow [3] and use their formulation. The problem can be restated with a single threshold parameter as given in the original paper [5].)

Max Margin (P1)

$$\begin{array}{lll} \text{min.} & \frac{1}{2}\|\boldsymbol{w}\|^2 - (\theta_+ - \theta_-) & \\ \text{w.r.t.} & \boldsymbol{w} = (w_1, ..., w_n), \theta_+, \text{ and } \theta_-, & \\ \text{s.t.} & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \geq \ \theta_+, & \text{if } y_i = 1, \\ & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \leq \ \theta_-, & \text{if } y_i = -1. \end{array}$$

In this formulation, it is assumed that the data set is linearly separable; that is, a hyperplane separating the two classes of examples exists. By the *solution* of (P1), we mean the hyperplane that achieves the minimum cost. Given a solution, its support vectors are the data points $\boldsymbol{x}_i$ for which, at the solution, the corresponding inequality is tight; that is, $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_+$ if $y_i = 1$, $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_-$ if $y_i = -1$.

An important feature of SVM is that it is also applicable for the nonseparable case. More precisely, for nonseparable data we can take two positions: (i) the case where we consider that a hyperplane is too weak to be a classifier for our given examples, and that we should be able to fit them better nonlinearly; and (ii) the case where we consider that there are some erroneous examples or exceptions, i.e., "outliers", which should be somehow identified and allowed to be misclassified. Of course, it would be better if we can use a "reasonable" nonlinear classifier. Nevertheless, the second approach is important as well if we suspect that outliers exist in a given set of examples. The usability of SVM is due to the fact that we can use both.

*Bioeng. Dept., Univ. Illinois at Chicago
†東京工業大学 社会理工学研究科 経営工学専攻
‡東京工業大学 情報理工学研究科 数理・計算科学専攻

In this paper, we focus on the second approach, and discuss how to deal with outliers. A standard SVM solution to the case where some outliers exist is to relax the constraints by introducing slack variables or "soft margin error". That is, we consider the following generalization of the problem (P1), corresponding to the soft margin hyperplane separation problem.

Max Soft Margin (P2)

$$\begin{array}{lll} \text{min.} & \frac{1}{2}\|\boldsymbol{w}\|^2 - (\theta_+ - \theta_-) + D\sum_i \xi_i & \\ \text{w.r.t.} & \boldsymbol{w} = (w_1, ..., w_n), \theta_+, \theta_-, & \\ & \text{and } \xi_1, ..., \xi_m, & \\ \text{s.t.} & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \geq \ \theta_+ - \xi_i, & \text{if } y_i = 1, \\ & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \leq \ \theta_- + \xi_i, & \text{if } y_i = -1, \\ & \text{and } \xi_i \ \geq \ 0. & \end{array}$$

For a given set X of examples, suppose we solve the problem (P2) and obtain the optimal hyperplane. Then an example in X is called an *outlier* if it is misclassified with this hyperplane. On the other hand, examples other than outliers are called *normal examples*. Notice that this definition of outlier is relative both to the hypothesis class and to the soft margin parameter D, which determines the degree of influence of the outliers. In this paper, we discuss a way to choose this parameter D appropriately.

Though important, there seems, at least as far as the authors know, no systematic method for choosing the influence parameter D. The most standard way [6] is to try several choices and use the one with the best performance on the training set. Note that D should be fixed in advance; that is, when solving (P2) (in other words, when training SVM), D is considered as a constant. But of course, we may be able to revise D depending the result of solving (P2). Thus, if there is a reasonable criterion for D depending on the solution of (P2), then we would at least have the following method to choose D: Solve (P2) by revising D until some D satisfying the criterion is obtained.

In this paper, we prove that D is small enough to satisfy a certain reasonable condition (see the next section) if and only if (P2) has a nontrivial solution. From this result, we propose to check whether D is appropriate by checking whether (P2) has a nontrivial solution.

2 Technical Discussion

Bennett and Bredensteiner [3] showed an alternative formulation of (P2), giving us an intuitive geometric interpretation of the influence parameter D. Our criterion is based on their geometric interpretation. So we start with a brief explanation of their formulation. For simplifying our discussion, let us assume in the following that

$D = 1/k$ for some integer. (That is, we will consider only those D's that are the inverse $1/k$ of some integer k.)

Note that each example $\boldsymbol{x}_i$ is a point in $\mathbb{R}^n$. Intuitively, the objective of (P1) is to find two parallel hyperplanes separating positive and negative points with the largest distance. This distance can be regarded as the distance between two sets of points, i.e., the sets of positive/negative points. Following the argument in [3], a similar interpretation is given to the formulation (P2). In the nonseparable case (i.e., in the formulation (P2)), we would not consider each example; instead, we would consider "composed examples" that are obtained as a center of k examples, and then consider the distance between the sets of positive/negative composed examples. Note that the number k is defined by $k = 1/D$.

For a more precise statement, consider the set Z of *composed examples* $\boldsymbol{z}_I$ that is defined by

$$\boldsymbol{z}_I \;=\; \frac{\boldsymbol{x}_{i_1} + \boldsymbol{x}_{i_2} + \cdots + \boldsymbol{x}_{i_k}}{k},$$

with some k distinct elements $\boldsymbol{x}_{i_1}$, $\boldsymbol{x}_{i_2}$, ..., $\boldsymbol{x}_{i_k}$ of X with the same label (i.e., $y_{i_1} = y_{i_2} = \cdots = y_{i_k}$). That is, a composed example is a mass center of all groups of k homogeneously labeled initial data points. The label y_I of the composed example $\boldsymbol{z}_I$ inherits its members'. In the following, we use I for indexing elements of Z and their labels. (For distinguishing from composed examples $\boldsymbol{z}_I$, we will call $\boldsymbol{x}_i$ an *original example*.)

By using the technique in [3], we can easily show that (P2) is essentially equivalent to the following (P5). (See the full version of this paper [2] for the precise meaning of "equivalent". We use (P5) in order to be consistent with [2].)

Max Margin for Composed Examples (P5)

$$\begin{array}{lll} \text{min.} & \frac{1}{2}\|\boldsymbol{w}\|^2 - (\eta_+ - \eta_-) & \\ \text{w.r.t.} & \boldsymbol{w} = (w_1, ..., w_n), \eta_+, \text{ and } \eta_-, & \\ \text{s.t.} & \boldsymbol{w} \cdot \boldsymbol{z}_I \;\geq\; \eta_+, & \text{if } y_I = 1, \\ & \boldsymbol{w} \cdot \boldsymbol{z}_I \;\leq\; \eta_-, & \text{if } y_I = -1. \end{array}$$

In general, composed examples may not be separable. But they are separable if k is sufficiently large (in other words, if $D = 1/k$ is sufficiently small). For example, in the extreme case where $k = m_+ = m_-$ (where m_+ and m_- are respectively the number of positive and negative examples), we have only one positive and one negative composed example, which are clearly separable (unless they are the same).

From this observation, one natural criterion for D (or k) is the linear separability of composed examples. Note, on the other hand, this criterion is rather difficult to use when solving (P2). Our technical contribution is to prove the following characterization theorem, thereby providing a way to determine the linear separability of composed examples. (See [2] for the proof.)

Theorem 2.1 *Let X be any set X of examples, and for any k, let Z be the set of composed examples made up from k examples from X. Let $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*, \boldsymbol{\xi}^*)$ be a solution of (P2) on X. Then $\|\boldsymbol{w}^*\| > 0$ (or, equivalently $\boldsymbol{w}^* \neq \boldsymbol{0}$) if and only if k is large enough so that Z is linearly separable.*

From this theorem, if D is too large and the composed examples are not linearly separable, then we know it by obtaining the trivial answer $\boldsymbol{w}^* = \boldsymbol{0}$ by solving (P2); on the other hand, if D is small enough (i.e., k is large enough) so that the composed examples are linearly separable, then (P2) should have a nontrivial answer. That is, we can determine whether D is appropriate by checking whether (P2) has a nontrivial solution.

In the context of linear programming type classification, Bennet and Mangasarian [4] proposed to use this "linear separability" for choosing weights that are similar to our influence parameter. There they also provide a way to decide whether examples are linearly separable under a given choice of weights. Here we prove that a similar characterization theorem.

It should be mentioned that our characterization theorem, though its proof is not so difficult, is not immediate from the equivalence between (P2) and (P5). This is because the equivalence is guaranteed *when* the composed examples are linearly separable.

3 Concluding Remarks

Assume, as our working hypothesis, that the separability of composed examples is an appropriate criterion for choosing the influence parameter D. Then from our theorem one can easily think of an algorithmic way to determine D under this criterion. The simplest and straightforward way is to use the binary search method to find the largest D $(= 1/k)$ such that (P2) has a nontrivial solution $\boldsymbol{w}^*$, i.e., $\boldsymbol{w}^* \neq \boldsymbol{0}$. For this, we need to solve (P2) several times with different D, which may be computationally hard if there are huge number of examples. We point out here that a random sampling approach proposed in [1] could be used to solve this hardness.

References

[1] J. L. Balcázar, Y. Dai, and O. Watanabe, Provably fast training algorithms for support vector machines, in *Proc. First IEEE International Conference on Data Mining* (ICDM'01), IEEE, 43−50, 2001.

[2] Y. Dai, J. Tanaka, and O. Watanabe, Research Report C-163 (2002). (Availabe from `www.is.titech.ac.jp/research-report/C/`.)

[3] K.P. Bennett and E.J. Bredensteiner, Duality and geometry in SVM classifiers, in *Proc. 17th Int'l Conf. on Machine Learning* (ICML'2000), 57−64, 2000.

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

[5] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* 20, 273−297, 1995.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press 2000.