Classification SVM algorithms with interval-valued training data using triangular and Epanechnikov kernels

Lev V. Utkin, Anatoly I. Chekh and Yulia A. Zhuk

Saint Petersburg State Forest Technical University

lev.utkin@gmail.com, anatoly.chekh@gmail.com, zhuk_yua@mail.ru

1. Classification problem by interval-valued training data

Given: training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, $y_i \in \{-1, 1\}$. $\mathbf{x}_i$ are interval-valued, i.e., $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$. Here $\mathbf{A}_i = [\underline{a}_i^{(1)}, \overline{a}_i^{(1)}] \times ... \times [\underline{a}_i^{(m)}, \overline{a}_i^{(m)}]$, i.e., $\underline{a}_i^{(k)} \leq x_i^{(k)} \leq \overline{a}_i^{(k)}$, $k = 1, ..., m$.

The learning problem: to select a function $f(\mathbf{x}, w_{\text{opt}})$ from a set of functions $f(\mathbf{x}, w)$, which separates examples of different classes y.

A general approach for the problem solution by precise data: to minimize the risk functional or expected risk:

$$R(\mathbf{w},b) = \int_{\mathbb{R}^m} l(\mathbf{w}, \phi(\mathbf{x}))\mathrm{d}F(\mathbf{x}),$$

with the loss function:

$$l(\mathbf{w}, \phi(\mathbf{x})) = \max\{0, b - \langle \mathbf{w}, \phi(\mathbf{x})\rangle\},$$

ϕ is a feature map $\mathbb{R}^m \rightarrow G$ into an alternative higher-dimensional feature space G.

The empirical expected risk

$$R_{\text{emp}}(\mathbf{w},b) = \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{w}, \phi(\mathbf{x}_i)).$$

1.1 Approaches for solving the problem by interval data

- Interval-valued data are replaced by precise values based on some assumptions, for example, by taking middle points of intervals (LimaNeto and Carvalho 2008): *a very popular approach, unjustified, especially, by large intervals*
- The standard interval analysis (Angulo 2008, Hao 2009): *only linear separating or regression functions*
- Bernstein bounding schemes (Bhadra et al. 2009): *incorporate probability distributions over intervals.*
- The Euclidean distance between two data points in the Gaussian kernel is replaced by the Hausdorff distance and other distances between two hyperrectangles (Do and Poulet 2005, Chavent 2006, Pedrycz et al 2008, Schollmeyer and Augustin 2013): *a nice and simple idea, but with some questions.*
- Minimizing and maximizing the risk over values of intervals (Utkin and Coolen 2011, Cattaneo and Wienzierz 2015): *only monotone separating functions* (Utkin and Coolen 2011) *or only interval-valued response variables y in regression models* (Cattaneo and Wienzierz 2015).

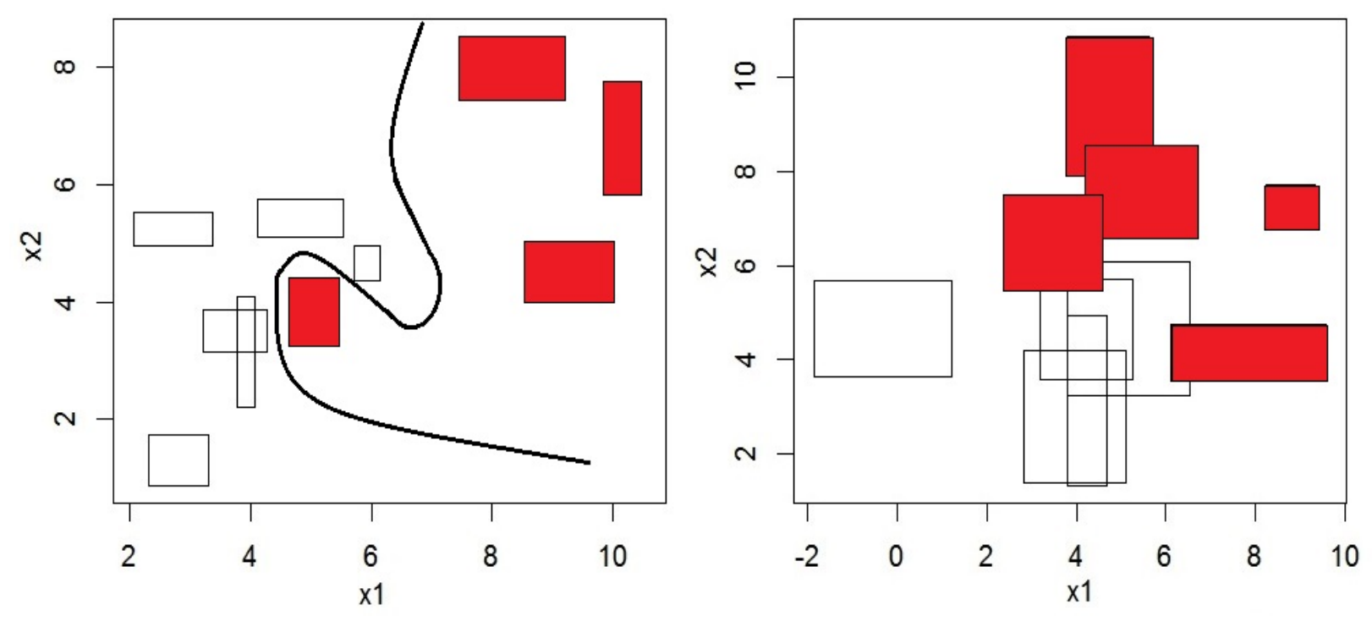


Figure 1: *Classification by small and large intervals*

1.2 Ideas underlying two new algorithms

1. Intervals produce a set of expected risk measures such that the upper risk is determined by maximizing the risk over values of intervals (*this is an old idea used in Utkin and Coolen 2011, Cattaneo and Wienzierz 2015*).
2. By applying the lower risk (**the minimax strategy**), it would be nice to isolate a "linear" program from the SVM with variables $\mathbf{x}_i \in \mathbf{A}_i$ and then to work with extreme points $\mathbf{x}_i^*$.
3. **Important idea**: We replace the Gaussian kernel by the triangular kernel (Utkin and Chekh 2015). This replacement allows us to get a set of linear programs with variables $\mathbf{x}_i$ restricted by $\mathbf{A}_i$, $i = 1, ..., n$.

1.3 The minimax strategy

A general approach for solution by interval data: to use the minimax strategy (Γ-minimax) and to minimize the upper $\overline{R}$ expectations of the loss function $l(\mathbf{x})$ in the framework of belief functions (Nguyen-Walker 1994, Strat 1990):

$$\begin{aligned} R(\mathbf{w}_{\text{opt}},b_{\text{opt}}) &= \min_{\mathbf{w},\rho} \overline{R}(\mathbf{w},b) \\ &= \sum_{i=1}^{n} m(\mathbf{A}_i) \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i). \end{aligned}$$

2. An algorithm with L_2-norm SVM

Support vector machine (SVM): a dual form (the Lagrangian)

$$\max_{\mathbf{x}_i \in \mathbf{A}_i} \max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, ..., n.$$

The separating function f in terms of Lagrange multipliers:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

We fix α and replace the Gaussian kernel by the triangular one:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2} \right) \rightarrow$$
$$\rightarrow T(\mathbf{x}, \mathbf{y}) = \max\left\{ 0, 1 - \frac{\|\mathbf{x} - \mathbf{y}\|^1}{\sigma^2} \right\}$$

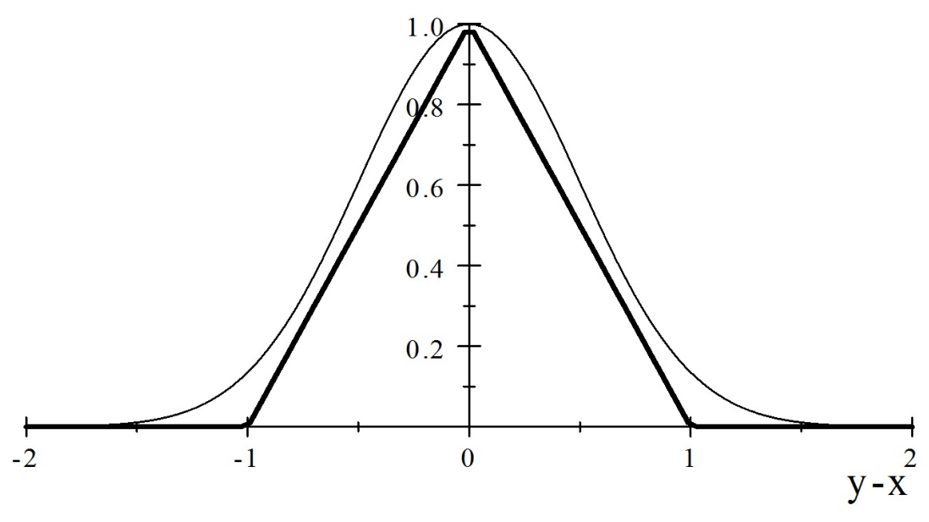


Figure 2: *The Gaussian and triangular kernels*

As a result, we get a set of standard quadratic problems:

- By fixed Lagrangian multipliers α and with the triangular kernel, we get a linear problem with constraints $\mathbf{x}_i \in \mathbf{A}_i$.
- Its optimal solution is achieved at extreme points of the hyper-rectangles produced by $\mathbf{A}_i$, i.e., at interval bounds.
- For every extreme point, we solve the standard quadratic problem.

The problem with absolute values:

Lemma 1 (Beaumont,1998) *If $[\underline{x}, \overline{x}] \subset \mathbb{R}$, $\underline{x} < \overline{x}$, and, if*

$$u = \frac{|\overline{x}| - |\underline{x}|}{\overline{x} - \underline{x}}, \quad v = \frac{\overline{x}|\underline{x}| - \underline{x}|\overline{x}|}{\overline{x} - \underline{x}},$$

we have

$$\forall x \in [\underline{x}, \overline{x}], \quad |x| \leq ux + v.$$

The main problem of the algorithm:

If we have n examples consisting of m features, then the number of extreme points (quadratic programs) is $t = 2^{nm}$.

3. An algorithm with L_infinite-norm SVM

Idea: There are many variants of SVMs, **it would be nice to find a SVM for which constraints for classification parameters do not depend on interval observations x_i.**

There is an interesting L_∞-norm SVM proposed by Zhou et al. 2002:

$$\min_{\alpha_j, \xi_j, j=1,...,n, r, b} R = \min_{\alpha_j, \xi_j, j=1,...,n, r, b} \left(-r + C\sum_{i=1}^{n} \xi_i \right),$$

subject to

$$y_j \left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq r - \xi_j, \ j = 1, ..., n,$$

$$-1 \leq \alpha_i \leq 1, \ i = 1, ..., n, \ \ r \geq 0, \ \xi_j \geq 0, \ j = 1, ..., n.$$

The dual form by fixed $\mathbf{x}_1, ..., \mathbf{x}_n$:

$$\min_{z} \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{n} z_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

subject to

$$\sum_{i=1}^{n} z_i \geq 1, \ 0 \leq z_j \leq C, \ j = 1, ..., n, \ \ \sum_{i=1}^{n} z_i y_i = 0.$$

All $\mathbf{x}_1, ..., \mathbf{x}_n$ are in the objective function. **Constraints have only variables $z_1, ..., z_n$ which produce the convex set Z of an interesting form.**

3.1 The convex sets of solutions

$$\textstyle\sum_{i=1}^{n} z_i \geq 1, \ 0 \leq z_j \leq C, \ j = 1, ..., n, \ \ \sum_{i=1}^{n} z_i y_i = 0.$$

$$z_1 \rightarrow y_1 = -1, \ z_2 \rightarrow y_2 = 1, \ z_3 \rightarrow y_3 = 1$$

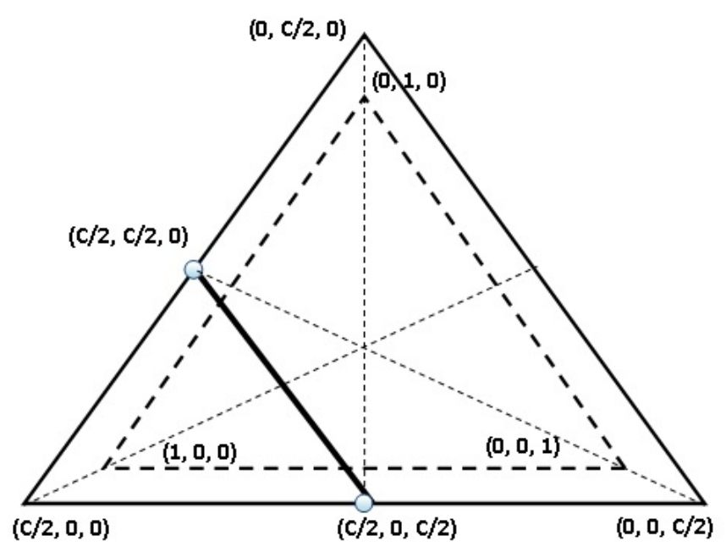


Figure 3: *The convex set Z*

Proposition 1 *Let $n_-$and n_+be numbers of $y = -1$and $y = 1$. $t$and s:*

$$(2C)^{-1} < t \leq \min(n_-, n_+),$$
$$(2C)^{-1} - 1 \leq s < \min\left((2C)^{-1}, n_-, n_+\right),$$

The first subset: $N_1 = \sum_{t=\lceil 1/2C \rceil}^{\min(n_-,n_+)} \binom{n_-}{t}\binom{n_+}{t}$extreme points: $t$elements from every class are C, others are 0. If $s \geq 0$, then the second subset: $N_2 = (n_- - s)(n_+ - s)\binom{n_-}{s}\binom{n_+}{s}$ extreme points: $s$elements from every class are C, one element from every class is $1/2 - sC$, others are 0.

The final optimization problems

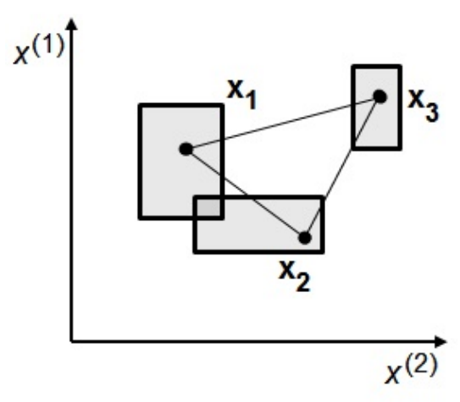
By using again the triangular kernel, we get a set of N_1+N_2 (the number of extreme points of Z) linear programs with variables $\mathbf{x}_i \in \mathbf{A}_i, i = 1, ..., n$. The number of linear program does not depend on the number m of features!

3.2 The "duality" of algorithms

- The first algorithm uses extreme points of interval-valued data $\mathbf{x}_i \in \mathbf{A}_i$ by fixed Lagrange multipliers α_i,
- The second algorithm uses extreme points of Lagrange multipliers z_i by fixed interval-valued data $\mathbf{x}_i \in \mathbf{A}_i$.

3.3 Precise values of intervals

The proposed algorithms produce unique and consistent precise points of intervals corresponding to the largest value of the expected classification risk.



4. The Epanechnikov kernel

Another kernel:

$$T_2(\mathbf{x}, \mathbf{y}) = \max\{0, 1 - \|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2\}.$$

We get a set of N_1+N_2 quadratically constrained linear programs (QCLP).

Tools: the sequential quadratic programming (Boggs and Tolle 1995), SNOP (Gill et al. 2002)

5. Advantages of the algorithms

1. The algorithms allows us to construct non-linear separating functions.
2. The algorithms are justified from the decision point of view (minimax strategy).
3. The algorithms produce unique precise points of intervals corresponding to the largest value of the expected risk. The points compose a single probability distribution among a set of distributions produced by intervals.
4. The algorithms can be extended on the support vector regression algorithms when dependent as well as independent variables are interval-valued.

ISIPTA 2015, 9th International Symposium on Imprecise Probability: Theory and Applications, July 2015, Pescara, Italy