New black-box modeling for novelty detection and semi-supervised learning

Vilen Jumutc Johan A.K. Suykens

Workshop on Data-driven Modeling
Methods and Applications, BIL 2014
Leuven, 14 – 15 July, 2014

Outline

1. Novelty Detection
2. Supervised Novelty Detection
3. Semi-Supervised Learning
4. Conclusion

Estimation of the support for a high-dimensional distribution

1. High-dimensional space might be difficult to interpret and understand
   - RBF and other non-linear kernels don't necessary help to reveal the underlying structure
   - Modelling of the support by well-known methods like One-Class SVM[1] explicitly presumes i.i.d assumption and a single high-dimensional distribution from which data points are drawn.

   Solution:
   - Break down the problem and model distributions separately.
2. i.i.d assumption might break
   - This means that we can face a mixture of distributions . . .
   - . . . and these distributions might interact.

   Solution:
   - We need to model these distributions explicitly.

[1] Schölkopf et al., Estimating the Support of a High-Dimensional Distribution, Neural Comput., vol 13(7), pp. 1443 – 1471, July 2001.

One-Class SVM

Primal form

$$\min_{w\in\mathcal{F};\xi\in\mathbb{R}^l;\rho\in\mathbb{R}} \|w\|^2 + \frac{1}{\nu l}\cdot\sum_l \xi_i - \rho \tag{1}$$

$$s.t. \qquad \langle w, \Phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i \tag{2}$$

Dual form

$$\max_{\alpha} -\alpha^T K \alpha, \tag{3}$$

$$s.t. \qquad \begin{array}{l} 0 \leq \alpha_i \leq \frac{1}{\nu l}, \ \forall i \\ \mathbf{1}^T\alpha = 1 \end{array} \tag{4}$$

Supervised Novelty Detection

1. Supervised Novelty Detection (SND) is designed for finding outliers in the presence of several classes.
2. SND estimates the support of the two separate classes (distributions) while trying to keep necessary discrimination between them.
3. SND supports the i.i.d. assumption and density estimation for all target classes (distributions) separately.
4. SND doesn't try to find outliers in the existing pool of training samples.

Potential applications

1. SND algorithm could be utilized in Intrusion Detection where several user groups are performing different actions on a system and our goal is to locate intruders among users.
2. SND can be utilized in bioinformatics where we have to locate misbehaving genes out of several possible groups in genome while keeping necessary discrimination along these groups.

A link to One-Class SVM and a new optimization problem

Primal form

$$\min_{w_1, w_2 \in \mathcal{F}; \xi, \xi^* \in \mathbb{R}^l; \rho_1, \rho_2 \in \mathbb{R}} \quad \frac{\gamma}{2}(\|w_1\|^2 + \|w_2\|^2) + \langle w_1, w_2 \rangle + C \cdot \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \rho_1 - \rho_2 \tag{5}$$

$$s.t. \quad \begin{aligned} & y_i(\langle w_1, \Phi(x_i) \rangle - \rho_1) + \xi_i \geq 0, && i \in \overline{1, l} \\ & y_i(\langle w_2, \Phi(x_i) \rangle - \rho_2) - \xi_i^* \leq 0, && i \in \overline{1, l} \\ & \xi_i \geq 0, \xi_i^* \geq 0, && i \in \overline{1, l} \end{aligned} \tag{6}$$

Decision functions

Decision functions

$$f_1(x) = \langle w_1, \Phi(x) \rangle - \rho_1, \tag{7}$$

$$f_2(x) = \langle w_2, \Phi(x) \rangle - \rho_2. \tag{8}$$

Decision rule

$$c(x) = \begin{cases} \text{argmax } \mathrm{f_{c_i}(x)}, & \text{if } \max \mathrm{f_{c_i}(x)} > 0 \\ c_{out}, & \text{otherwise} \end{cases} \tag{9}$$

Lagrangian

Dual form

$$\mathcal{L}_D(\alpha, \lambda) = \frac{\mu_1}{2}(\alpha^T G \alpha + \lambda^T G \lambda) + \mu_2(\alpha^T G \lambda), \tag{10}$$

$$\begin{aligned} & G = K \circ yy^T, \\ & \mu_1 = \tfrac{\gamma}{1-\gamma^2}, \\ s.t. \quad & \mu_2 = \tfrac{1}{1-\gamma^2}, \\ & C \geq \alpha_i \geq 0,\ C \geq \lambda_i \geq 0,\ \forall i \\ & y^T \alpha = 1,\ y^T \lambda = -1 \end{aligned} \tag{11}$$

Decision functions in terms of dual variables

Decision functions

$$f_1(x) = \frac{\gamma \sum_{i=1}^{l} \alpha_i y_i k(x_i, x) + \sum_{i=1}^{l} \lambda_i y_i k(x_i, x)}{\gamma^2 - 1} - \rho_1, \qquad (12)$$

$$f_2(x) = \frac{\gamma \sum_{i=1}^{l} \lambda_i y_i k(x_i, x) + \sum_{i=1}^{l} \alpha_i y_i k(x_i, x)}{1 - \gamma^2} - \rho_2. \qquad (13)$$

The latter functions we can plug-in into Eq.(9) and get the evaluation of any out-of-sample data point w.r.t. SND model.

Extension to multi-class setting

Dual form

$$\max_{\alpha_i} \mathcal{L}_D(\alpha_i) = \frac{1}{\mu} \sum_{i}^{n_c} \lambda_i^T K \alpha_i, \tag{14}$$

$$\text{s.t.} \quad \begin{array}{l} C \geq \alpha_{ij} \geq 0, \ \forall i \in \overline{1, n}, \ \forall j \in \overline{1, n_c} \\ \sum_{i=1}^{n} \alpha_{ij} = 1, \ \forall j \in \overline{1, n_c} \end{array} \tag{15}$$

where $\lambda_i = (\gamma + n - 2)(\alpha_i \circ y_i) - \sum_{j=1, j \neq i}^{n_c} (\alpha_j \circ y_j)$, $\mu = (\eta + 1)(\gamma - 1)$, K is a kernel matrix of size $n \times n$, $\circ$ denotes component-wise multiplication and n_c is the number of classes.

Least-Squares SND

Linear system

$$\left(\begin{array}{c|c} \begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix} & \begin{matrix} 1_1^T & \dots & 0_n^T \\ \vdots & \ddots & \vdots \\ 0_n^T & \dots & 1_n^T \end{matrix} \\ \hline \begin{matrix} 1_n & \dots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \dots & 1_n \end{matrix} & \begin{matrix} -\frac{\eta G^\star_{11}}{\mu} & \dots & \frac{G_{1n_c}}{\mu} \\ \vdots & \ddots & \vdots \\ \frac{G_{n_c1}}{\mu} & \dots & -\frac{\eta G^\star_{n_cn_c}}{\mu} \end{matrix} \end{array}\right) \left(\begin{array}{c} \rho_1 \\ \vdots \\ \rho_{n_c} \\ \hline \alpha_1 \\ \vdots \\ \alpha_{n_c} \end{array}\right) = \left(\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline 0_n \\ \vdots \\ 0_n \end{array}\right) \quad (16)$$

where $G^\star_{ij} = G_{ij} + \frac{\mu}{\eta\gamma_2} I$, 1_n and 0_n denote vectors of length n

UCI datasets and evaluation

Table : Averaged misclassification error

Dataset	SND	*C*-SVM	LS-SVM
Toy Data (1)	0.1395± 0.097	0.1385± 0.078	**0.1325**± 0.085
Arcene	**0.1620**± 0.006	0.1730± 0.095	0.1810± 0.091
Ionosphere	0.0684± 0.043	0.0740± 0.031	**0.0483**± 0.030
Parkinsons	**0.0613**± 0.046	0.0721± 0.060	0.0621± 0.064
Sonar	**0.0962**± 0.069	0.1250± 0.105	0.1205± 0.101
Zoo	**0.0500**± 0.081	0.0733± 0.119	0.1071± 0.119
Iris	0.0467± 0.068	**0.0440**± 0.065	0.0493± 0.067
Ecoli	0.1263± 0.069	**0.1240**± 0.061	0.1562± 0.062
TAE	**0.4031**± 0.159	0.4346± 0.146	0.5545± 0.131
Seeds	0.0667± 0.060	**0.0650**± 0.050	0.0838± 0.073

Novelty detection experiments

Table : Averaged misclassification/detection rate

Dataset	SND	One-Class SVM
Toy Data (2)	**0.0083** / (0.9746)	0.0233 / (1)
Toy Data (3)	**0.0113** / (1)	0.0233 / (1)
Toy Data (4)	**0.0366** / (0.8182)	0.0791 / (0.7808)

Toy problems explained

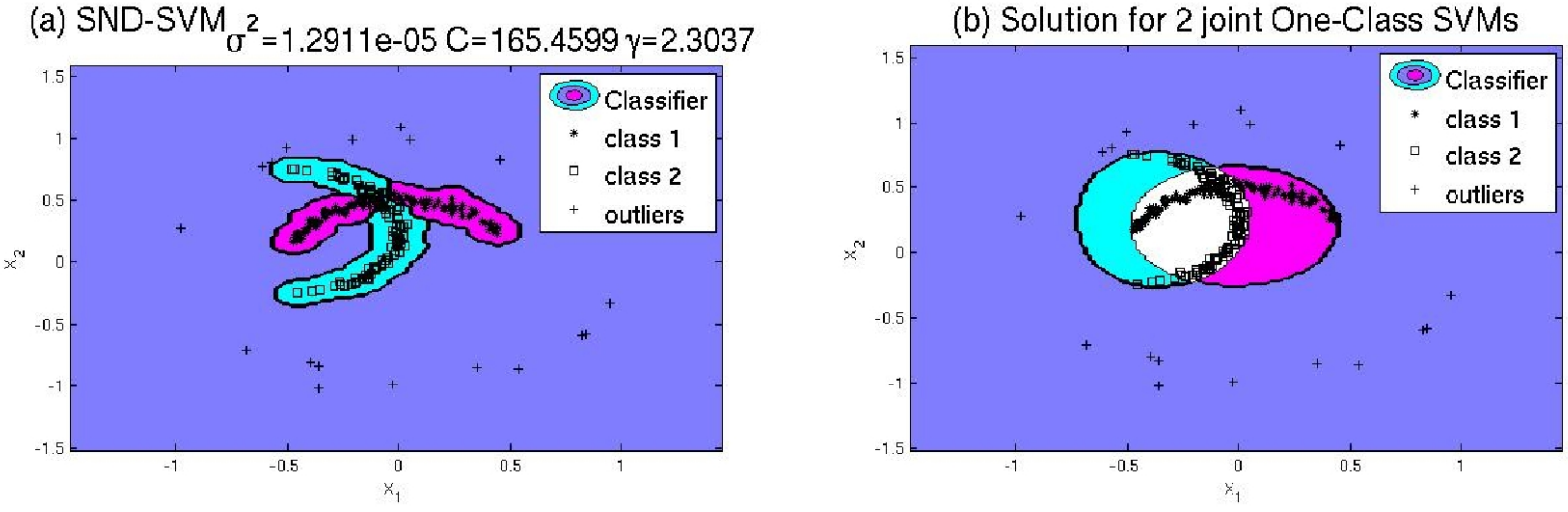

Figure : Comparison of SND (a) and two joint One-Class SVMs (b) in the novelty detection scheme showing a clear improvement of SND. White region depicts the area which belongs to the support of both One-Class SVMs simultaneously.

Bilinear Semi-Supervised KSC

We try to apply the same reasoning *w.r.t.* semi-supervised learning and incorporate widely acknowledged Kernel Spectral Clustering (KSC) model

Primal form

$$\begin{aligned} \min_{w_1,w_2\in\mathbb{R}^d;e_1,e_2\in\mathbb{R}^m} \quad & \tfrac{\gamma_1}{2}(\|w_1\|^2+\|w_2\|^2)+\langle w_1,w_2\rangle \\ & +\tfrac{\gamma_2}{2}\textstyle\sum_{i=1}^{l}[(e_{1i}-y_i)^2+(e_{2i}+y_i)^2] \\ & -\tfrac{\gamma_3}{2}e_1^TVe_1-\tfrac{\gamma_4}{2}e_2^TVe_2 \\ \text{s.t.} \quad & \Phi w_1=e_1 \\ & \Phi w_2=e_2 \end{aligned} \tag{17}$$

where $w_1, w_2 \in \mathbb{R}^h$ are the hyperplanes associated with each non-parallel classifier, $\Phi = [\varphi(x_1)^T;\ldots;\varphi(x_m)^T] \in \mathbb{R}^{m\times h}$ is a stacked feature map for all samples in $\mathcal{X}$, $e_1, e_2 \in \mathbb{R}^m$ are the projections, V is the weighting diagonal matrix and γ_1 through γ_4 denote trade-off hyperparameters

Linear system to solve

Linear system

$$\left[\begin{array}{c|c} A_d & A_{\text{off}} \\ \hline L_{\text{off}} & L_d \end{array}\right] \left[\frac{\alpha}{\lambda}\right] = \gamma_2 \left[\frac{y^*}{-y^*}\right], \tag{18}$$

where $A_d = \gamma_3\gamma_1\gamma_1'\Omega D^{-1} - \gamma_2\gamma_1\gamma_1' G + I$,
$L_d = \gamma_4\gamma_1\gamma_1'\Omega D^{-1} - \gamma_2\gamma_1\gamma_1' G + I$, $L_{\text{off}} = \gamma_2\gamma_1'\Omega - \gamma_4\gamma_1' G D^{-1}$,
$A_{\text{off}} = \gamma_2\gamma_1' G - \gamma_3\gamma_1'\Omega D^{-1}$, $\gamma_1' = 1/(1-\gamma_1^2)$, Ω is a full-rank kernel matrix and G, D^{-1} were given in the previous subsection, while I is an identity matrix of size $m \times m$.

Evaluation

Table : Averaged misclassification for the benchmark SSL datasets

Dataset	Bi-SemiKSC	LapSVM	SemiKSC
g241c	**0.148**$\pm$0.044	0.232$\pm$0.052	0.194$\pm$0.102
g241d	**0.158**$\pm$0.034	0.278$\pm$0.062	0.191$\pm$0.069
BCI	**0.244**$\pm$0.056	0.376$\pm$0.049	0.315$\pm$0.078
Text	**0.382**$\pm$0.082	0.458$\pm$0.026	0.477$\pm$0.062

Conclusion

1. Algorithms incorporating bilinear term can estimate the support of the two separate distributions under separate i.i.d. assumptions.
2. Both SND and Bi-SemiKSC algorithms utilize information about labeling.
3. Algorithms demonstrate comparable or even smaller generalization errors.
4. Algorithms capture underlying structure more precisely than One-Class SVM or other semi-supervised learning techniques.

Thank you for attention!