ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Converting SVDD Scores into Probability Estimates

Meriem El Azami[1], Carole Lartizien[1] and Stéphane Canu[2] *

1- Université de Lyon, CREATIS; CNRS UMR5220; Inserm U1044;
INSA-Lyon; Univ. Lyon 1 - France
2- LITIS, INSA de Rouen, Normandie Université,
Saint-Etienne-du-Rouvray, 76801, France

Abstract. To enable post-processing, the output of a support vector data description (SVDD) should be a calibrated probability as done for SVM. Standard SVDD does not provide such probabilities. To create probabilities, we first generalize the SVDD model and propose two calibration functions. The first one uses a sigmoid model and the other one is based on a generalized extreme distribution model. To estimate calibration parameters, we use the consistency property of the estimator associated with a single SVDD model. A synthetic dataset and datasets from the UCI repository are used to compare the performance against a robust kernel density estimator.

1 Introduction

Support vector classification methods, such as support vector data description (SVDD) [1] and its variants [2], have been successfully applied in the context of outlier detection [3]. The outputted scores however are very hard to interpret. In many application domains, and especially in pattern recognition methods, transforming these scores into well calibrated probabilities can help greatly with 1) score interpretation 2) threshold selection, and 3) score combination (e.g. ensemble learning methods). In [4], the decision boundary computed by OC-SVM was proven to converge to the minimum volume set (MV-set) with probability mass of at least $\eta = 1 - \nu$. Various attempts have been made to convert outlier scores into calibrated probabilities [5, 6]. All these methods however either make some assumption on the distribution of the outlier scores and no calibration scheme was specifically designed to convert the output of SVDD.

In the present study, our goal is 1) to propose a generalization of the SVDD method that allows estimating q MV-sets of given probability masses $\eta_j, j = 1 \dots q$, 2) convert the outputted outlier scores into probability estimates and 3) maximize the detection rate. The proposed approach is evaluated in the context of outlier detection and compared against the robust version of kernel density estimator (rKDE) that was recently proposed by Kim and Scott [7].

*This work was supported by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

2 SVDD generalization

2.1 Naive approach (iSVDD)

We first propose a straightforward generalization of the SVDD algorithm for the estimation of q MV-sets by constructing q independent SVDD models. For each $j = 1, \ldots, q$, the SVDD algorithm tries to find the enclosing hyper-sphere, of centre a_j and radius R_j, with minimum volume given ν_j. Let $(\mathbf{x_1}, \ldots, \mathbf{x_n})$, $\mathbf{x_i} \in \mathbb{R}^p$ be n training samples from a normal class (N). The global ν-formulation for all q SVDD models is:

$$\begin{cases} \min\limits_{R_j, \mathbf{a_j}, \xi_\mathbf{j}} & \sum\limits_{j=1}^{q} R_j^2 + \sum\limits_{j=1}^{q} \dfrac{1}{\nu_j n} \sum\limits_{i=1}^{n} \xi_{ji} \\ \text{s.t} & (\mathbf{x_i} - \mathbf{a_j})^\top (\mathbf{x_i} - \mathbf{a_j}) \leq R_j^2 + \xi_{ji}, \quad i = 1, \ldots, n \ , \ j = 1, \ldots, q \\ \text{and} & \xi_{ji} \geq 0, \qquad\qquad\qquad\qquad\quad i = 1, \ldots, n \ , \ j = 1, \ldots, q, \end{cases} \tag{1}$$

where ξ_{ji} are slack variables that allow relaxing the inequality constraints and $\nu_j \in [0, 1]$ corresponds to an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors [2]. The outlier score associated to a given observation $\mathbf{x}$ by the j^{th} independent SVDD model is then given by: $g_j(\mathbf{x}) = (\mathbf{x} - \mathbf{a_j})^\top (\mathbf{x} - \mathbf{a_j}) - R_j^2$. In [4], the consistency property of the SVDD algorithm was proven for estimating MV-sets. An estimate of the level-set of probability mass at least $1 - \nu_j$ is given by: $MV_{1-\nu_j} = \{\mathbf{x} \mid g_j(\mathbf{x}) = 0\}$.

2.2 Concentric SVDD models (cSVDD)

We extend the SVDD algorithm by constructing q hierarchical MV-sets with decreasing probability masses. This translates into having the same centre $\mathbf{a}$ for all q SVDD models and solving a single optimization problem. Let R_j be the radius associated with the j^{th} SVDD model, the primal problem formulation is:

$$\begin{cases} \min\limits_{R_j, \mathbf{a}, \xi_\mathbf{j}} & \sum\limits_{j=1}^{q} R_j^2 + \sum\limits_{j=1}^{q} \dfrac{1}{\nu_j n} \sum\limits_{i=1}^{n} \xi_{ji} \\ \text{s.t} & (\mathbf{x_i} - \mathbf{a})^\top (\mathbf{x_i} - \mathbf{a}) \leq R_j^2 + \xi_{ji}, \quad i = 1, \ldots, n \ , \ j = 1, \ldots, q \\ \text{and} & \xi_{ji} \geq 0, \qquad\qquad\qquad\quad i = 1, \ldots, n \ , \ j = 1, \ldots, q. \end{cases} \tag{2}$$

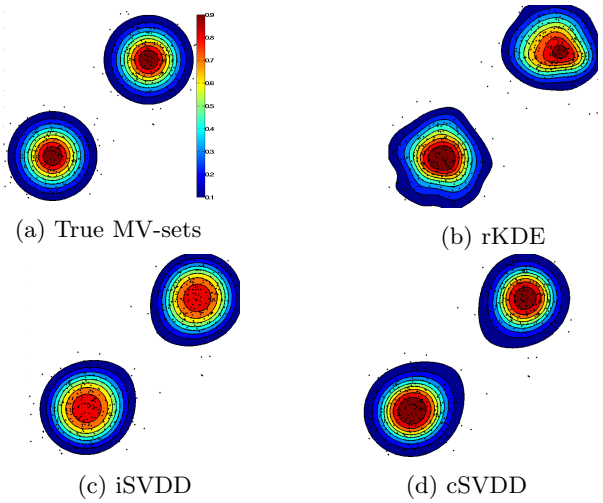
The dual formulation of problem 2 can be derived by considering the Lagrange multipliers and setting Karush-Kuhn-Tucker optimality conditions. The resulting dual formulation is very similar to that of the standard SVDD and the kernelization of this algorithm is straightforward by using the kernel trick and associated representer theorem.

The outlier score assigned to a given observation $\mathbf{x}$ by the j^{th} SVDD model is then given by : $f_j(\mathbf{x}) = f_c(\mathbf{x}) - R_j^2$ where $f_c(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^\top (\mathbf{x} - \mathbf{a})$. The center $\mathbf{a}$ is given by the representer theorem above. The different radii $R_j, \ j \in [1 \ldots q]$ can be obtained by considering for instance the distance to the center $\mathbf{a}$ of the essential support vectors for the j^{th} model.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

2.3 Method comparison

Figure 1 illustrates the estimated MV-sets for a bimodal Gaussian distribution. The true MV-sets 1a were computed using Monte Carlo simulation. For all methods, the kernel width was optimized to obtain the best estimation of the true MV-sets. Unlike rKDE, iSVDD and cSVDD both succeeded in capturing the nested nature of the MV-sets. cSVDD allowed a better estimate than iSVDD of the MV-set with the highest probability ($p = 0.9$) of belonging to the distribution.



(a) True MV-sets (b) rKDE

(c) iSVDD (d) cSVDD

Fig. 1: Minimum volume sets obtained for a bimodal Gaussian distribution using: (b) a robust kernel density estimator and (c-d) 2 SVDD generalizations. The black dots represent the training data ($n = 500$).

3 Score conversion into probabilities

Converting outliers scores into calibrated probabilities is a challenging task. Standard calibration methods such as in [8] and included references cannot be used since no labelled example is available for the outlier class. We propose two methods to model the distribution of outlier scores: fitting a sigmoid calibration function inspired by [9] or a generalized extreme value distribution.

3.1 Calibration using sigmoid function (sig)

Let f_i, $i \in [1, n]$ be outlier scores assigned to $\mathbf{x_i}$ by either cSVDD ($f_i = f_{c_i}(\mathbf{x_i})$) or iSVDD ($f_i = g_{0_i}(\mathbf{x_i})$) with f_{c_i} and g_{0_i} as defined in 2. Following [9], we propose to model the probability that $\mathbf{x_i}$ is an outlier given its outlier score $p(\mathbf{x_i}) = P(O|f_i)$, where O is the outlier class, as a decreasing sigmoid function:

$$p(\mathbf{x_i}) = \frac{A}{1 + exp(Bf_i + C)},$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

where A, B and C are the 3 model parameters that are estimated using the available level-sets. Indeed, the decision boundary of each SVDD model j is an estimate of the MV-set associated with a probability mass η_j that can be approximated by $1 - \nu_j$ [4]. Thus, an estimate of the probability $p(\mathbf{x_i})$ of any observation $\mathbf{x_i}$ lying on the decision boundary of the j^{th} SVDD model is known and is given by ν_j. For cSVDD, we directly construct the training set $\{(p_j, f_j), j = 1 \dots q\}$ by associating the cSVDD score f_j for all support vectors lying on the decision boundary of the j^{th} SVDD model to the corresponding ν_j value. For iSVDD, we select a given reference model j_0 and compute the scores corresponding to the essential support vectors of the other remaining SVDD models (i.e. the observations lying on the decision boundary of each remaining model) to form a training set composed of q coupled values (p_j, f_j).

3.2 Calibration using extreme value distributions (gev)

Our initial assumption is that outliers correspond to observations with very low occurrence probability and are located in the tail of the distribution. We are therefore interested in estimators that allow a good estimation of the tail of the unknown distribution $\mathbb{P}$. Extreme value theory is a branch of statistics that deals with extreme deviations of a probability distribution [10]. The extreme value probability (EVP) of a random variable z is the probability of z being the largest value of the dataset. A key theorem in extreme value statistics states that the EVP distribution can be expressed as the generalized extreme value (GEV) distribution. Its cumulative distribution is given by:

$$F(z) = exp(-[1 + \zeta(\frac{z-\mu}{\sigma})]^{-\frac{1}{\zeta}}),$$

where the parameter ζ is the shape parameter that controls the tail behaviour of the distribution. To use EVP as an outlierness measure, we consider the univariate distribution of scores f_i over the entire training dataset to fit the GEV distribution.

4 Experiments

4.1 Evaluation and parameter selection

For synthetic datasets, the true probability distribution is known, we therefore used the Kullback-Leibler divergence to measure the performance. For real dataset, we used the performance measure introduced in [6] to evaluate the impact of using calibrated scores on performance while taking into account both the reliability of the probability estimates and each class cardinality.

$$\text{Error Cost} = \frac{1}{2}\sum_{x \in N} P(O|x) \times \frac{1}{|N|} + \frac{1}{2}\sum_{x \in O} P(N|x) \times \frac{1}{|O|}.$$

Parameter tuning for outlier detection method is very hard as no labelled example from the outlier class is available during the training phase. We propose

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Dataset	Noise ratio 0%				Noise ratio 5%			
	iSVDD	cSVDD gev	cSVDD sig	rKDE	iSVDD	cSVDD gev	cSVDD sig	rKDE
Blood transfusion	0.49	**0.46**	0.48	**0.46**	0.26	0.26	**0.25**	0.33
Breast cancer	**0.25**	0.26	**0.25**	0.33	**0.23**	0.25	0.25	0.29
SPECTF heart	0.60	**0.54**	0.60	0.61	0.61	**0.54**	0.60	0.67
Banana	0.42	0.28	**0.27**	0.36	0.43	0.30	**0.28**	0.38
Balance LB vs R	0.33	0.32	**0.30**	0.33	0.35	0.36	0.36	**0.33**
Balance RB vs L	0.30	**0.27**	**0.27**	0.32	0.31	**0.28**	**0.28**	0.33
Pima indian	**0.38**	0.39	0.39	0.40	0.39	**0.38**	0.39	0.39

Table 1: Error cost for all 3 SVDD generalizations and rKDE.

to take advantage of the probabilistic interpretation of the calibrated SVDD outlier scores to tune the model parameters. The quality of the estimated MV-sets is measured using the relative difference $\frac{\eta_j - \tilde{\eta}_j}{\eta_j}$ between the expected probability mass (η_j) of each estimated MV-set and the experimental probability mass ($\tilde{\eta}_j$) computed on a validation set.

4.2 Experimental Results

Gaussian distribution: The training data consists of $N = 500$ data points from a 2D Gaussian distribution $N(0, 1)$. The true probability function is given by the χ_2 cumulative distribution function with 2 degrees of freedom. To simulate the presence of noise, outliers drawn from a uniform distribution $U_{[-5,5]}$ were added to the training data. The noise ratio was varied between 1% and 10%. All experiments on synthetic data were repeated 100 times. For both SVDD generalizations we used a linear kernel. For rKDE, the best results were obtained

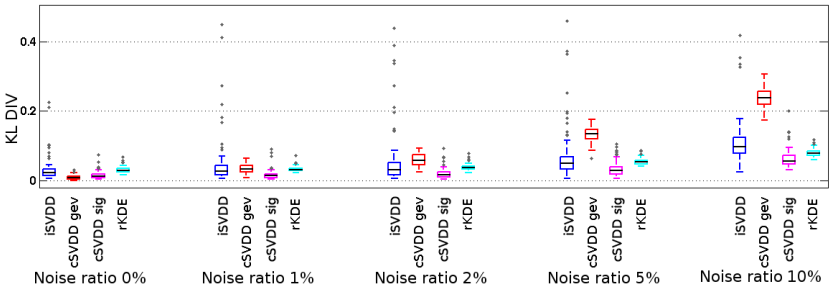


Fig. 2: KL divergence computed using the true probability and scores obtained with iSVDD (blue), cSVDD-gev (red), cSVDD-sig (magenta) and rKDE (cyan) for varying noise levels.

when using Hampel loss and least square cross validation to select the kernel bandwidth. Figure 2 shows that for all non null noise levels, the cSVDD-sig performs best and is the least sensitive to the presence of noise in the training data. The iSVDD approach seems to be less stable and generates more degenerate models than all other approaches.

Real datasets: we used six benchmark datasets from the UCI repository. For

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

each dataset, 200 examples were sampled from the majority class to form the normal class training examples. The second class was sampled to obtain the outliers. The Balance dataset contains three classes (L, B and R), we therefore trained using the two classes that allow having more than 200 instances and pooled the two remaining categories to form the outlier class. The presence of noise was simulated by adding a portion of examples from the outlier class to the training data. Table 1 shows that, for most considered dataset, the proposed generalizations along with the calibration give improved performance in terms of quality of the probability estimates and detection performance.

5 Conclusion

We introduced a generalization of the SVDD approach for estimating hierarchical MV-sets with specific probability masses. This generalization is obtained by transforming the initial SVDD optimization problem to construct q SVDD models that all share the same center. Given the consistency of the MV-sets estimated by each SVDD model, two calibration functions, a sigmoid model and a generalized extreme value distribution, have also been proposed to convert the outputted scores into calibrated probability estimates. Parameter tuning was automatically achieved by considering an optimality criterion evaluating probability estimates. The criterion does not require having labelled observations from the outlier class and makes no assumption on the distribution of outliers.

References

[1] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[2] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection. *ACM Computing Surveys*, 41(3):1–58, 2009.

[4] Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *JMLR*, 7:817–854, 2006.

[5] Jing Gao and Pang Ning Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 212–221, 2006.

[6] HP Kriegel, P Kröger, Erich Schubert, and Arthur Zimek. Interpreting and Unifying Outlier Scores. *SDM*, pages 13–24, 2011.

[7] JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, pages 2529–2565, 2012.

[8] Vojtech Franc, Alexander Zien, and Bernhard Schölkopf. Support vector machines as probabilistic models. In *ICML 2011*, pages 665–672, 2011.

[9] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[10] James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.