Drug Activity Characterization Using One-Class Support Vector Machines with Counterexamples

Alicia Hurtado-Cortegana[1], Francesc J. Ferri[1,⋆], Wladimiro Diaz-Villanueva[1], and Carlos Morell[2]

[1] Departament d'Informàtica, Universitat de València, Spain
[2] Comp. Sci. Dept. Univ. Central "Marta Abreu" de Las Villas, Santa Clara, Cuba

Abstract. The problem of detecting chemical activity in drugs from its molecular description constitutes a challenging and hard learning task. The corresponding prediction problem can be tackled either as a binary classification problem (active versus inactive compounds) or as a one class problem. The first option leads usually to better prediction results when measured over small and fixed databases while the second could potentially lead to a much better characterization of the active class which could be more important in more realistic settings. In this paper, a comparison of these two options is presented when support vector models are used as predictors.

1 Introduction

Among supervised learning techniques developed and widely used in recent years, support vector machines (SVM) have received considerable attention due both to their success in solving practical problems and their mathematical soundness. One of the distinguishing trends of SVM is their capability of generalization in the context of hard learning problems. Consequently, the literature exhibits lots of classification, clustering or regression problems spanning diverse application domains that can be very conveniently solved using SVM [1,2,3].

Data domain description, also referred to as one-class classification (OCC) constitutes a different prediction task which consists of characterizing only one class of objects (and consequently rejecting the rest). Depending on how the problem is posed, the differences with regard to two-class classification can be very subtle. The most important difference is that OCC aims at modeling a particular class instead of separating objects from two classes which implies modeling their discriminating boundary. One of the main consequences is the way in which both approaches treat outliers and novelties [4].

OCC models can be learned either from examples only or both from examples and counterexamples. In any case, the problem consists of arriving at a decision function that covers all examples without including any other regions in the representation space and excluding also all counterexamples, if any.

[⋆] Work partially funded by FEDER and Spanish Government through projects TIN2009-14205-C04-03 and Consolider Ingenio 2010 CSD07-00018.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 33–40, 2013.

In the particular case of support vector based approaches, several formulations exist. In particular, One-Class Support Vector Machines (OC-SVM) [5] try to learn a hyperplane in the Reproducing Kernel Hilbert Space (RKHS) that keeps examples as far as possible from the origin. On the other hand, Support Vector Data Description (SVDD) [4] consists of obtaining a kernelized hypersphere that contains all examples. These two formulations have been shown to be equivalent under certain circumstances [6]. In a recent work, SVDD has been extended by introducing a separation margin between examples and counterexamples [7]. In this way, the model not only optimally represents the class of interest but also robustly separates both types of data at the same time.

The purpose of the present work is to study advanced OCC models on a particular difficult task in which binary SVM arrive at very good solutions. The goal consists of assessing possible benefits and disadvantages of using more complex models to solve these challenging problems.

2 Learning Problem Formulations

Only the SVDD formulation and extensions are to be considered in the present work. Assume that data belong to a d-dimensional vector space, $\mathbb{R}^d$. There is also a mapping ϕ, from $\mathbb{R}^d$ to a RKHS, $\mathcal{H}$, which is implicitly given by a Mercer kernel function, $k : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^{\geq 0}$ in such a way that $k(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$, where $\langle \cdot,\cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$ [6] .

Let us suppose we have a non empty positive training set given by the examples corresponding to the class of interest, $\mathcal{X}^+ = \{x_1, \ldots, x_{\ell_1}\}$ and a negative training set which consists of zero or more counterexamples, $\mathcal{X}^- = \{x_{\ell_1+1}, \ldots, x_{\ell_1+\ell_2}\}$. The size of the overall training set, $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^-$, is given by $\ell = \ell_1 + \ell_2$. Each object from $\mathcal{X}$ has a corresponding label, y_i such that $y_i = 1$ if $1 \leq i \leq \ell_1$ and $y_i = -1$ if $\ell_1 < i \leq \ell$.

When only positive examples are to be used, SVDD tries to enclose all objects into a minimal hypersphere in the RKHS [4]. The so-called soft formulation introduces additional slack variables controlled by a penalty term to allow objects outside the hypersphere.

The formulation of the problem using a ν parameter is

$$\min_{R,c,\xi} R^2 + \frac{1}{\nu \ell_1} \sum_{i=1}^{\ell_1} \xi_i, \tag{1}$$

$$\text{subject to:} \quad \left(\|\phi\,(x_i) - c\|^2 - R^2 \right) \leq \xi_i, \tag{2}$$

$$\xi_i \geq 0. \tag{3}$$

By introducing a Lagrange multiplier, α_i, for each constraint it is possible to go from this primal formulation to its corresponding dual in which the optimization is over a vector, $\alpha = (\alpha_1, \ldots, \alpha_{\ell_1})$, which consists of all Lagrange multipliers in the primal problem.

$$\max_{\alpha} \sum_{i=1}^{\ell_1} \alpha_i k\left(x_i, x_i\right) - \sum_{i,j=1}^{\ell_1} \alpha_i \alpha_j k\left(x_i, x_j\right), \tag{4}$$

$$\begin{gathered}\text{subject to } 0 \leq \alpha_i \leq \frac{1}{\nu \ell_1}, \\ \sum_i \alpha_i = 1. \end{gathered} \tag{5}$$

This quadratic problem can be solved using the same methods as for binary SVMs. Once α has been obtained, the center of the hypersphere, c can be obtained from the additional constraint $c = \sum_{i=1}^{\ell_1} \alpha_i \phi(x_i)$. Correspondingly, the radius, R, can be obtained exactly in the same way as the bias of the linear function is computed in the case of binary SVMs [5]. The final characterization of the positive class is then given by the following decision function

$$f\left(x\right) = sgn\left(R^2 - \left\|\phi\left(x\right) - c\right\|^2\right) \tag{6}$$

The basic approach can be extended by introducing negative objects (counterexamples) and the corresponding constraints that keep them outside the hypersphere [8]. This introduces a sign (the label y_i) in the constraints and a new summation term in Eq. 1. Both summation terms will be now weighted by $\frac{\gamma}{\nu\ell}$ and $\frac{1-\gamma}{\nu\ell}$, respectively. γ is a new parameter that controls the relative importance of the constraint violations in both positive and negative cases which may be very important in specific practical problems exhibiting some kind of imbalance.

Apart from keeping positive data inside the hypersphere and negative data outside, it is possible to impose a (maximal) margin between the negative objects and the boundary of the hypersphere. This is the rationale of the Small Sphere and Large Margin (SSLM) approach [7]. The formulation of the corresponding primal problem in our particular context is:

$$\min_{R,c,\rho,\xi} R^2 - \eta\rho^2 + \frac{\gamma}{\nu\ell} \sum_{i=1}^{\ell_1} \xi_i + \frac{1-\gamma}{\nu\ell} \sum_{i=\ell_1+1}^{\ell} \xi_i \tag{7}$$

$$\begin{aligned} \text{subject to } & \left\|\phi\left(x_i\right) - c\right\|^2 \leq R^2 + \xi_i, && 1 \leq i \leq \ell_1 \\ & \left\|\phi\left(x_i\right) - c\right\|^2 \geq R^2 + \rho^2 - \xi_i, && \ell_1 < i \leq \ell \\ & \xi_i \geq 0, && 1 \leq i \leq \ell \end{aligned} \tag{8}$$

In this extended formulation, apart from the parameter ν that controls how strict the characterization must be, and the parameter γ that controls the trade off between positive and negative outliers, a new parameter η that moderates the maximization of the margin has been introduced. The margin is represented by a new variable, ρ.

These three OCC models constitute a family of predictors with increasing level of complexity. The more complex models need more parameters and the corresponding tuning gets harder. On the other hand, the more complex models are able to attain better characterizations with improved separation which will potentially lead to better generalization abilities.

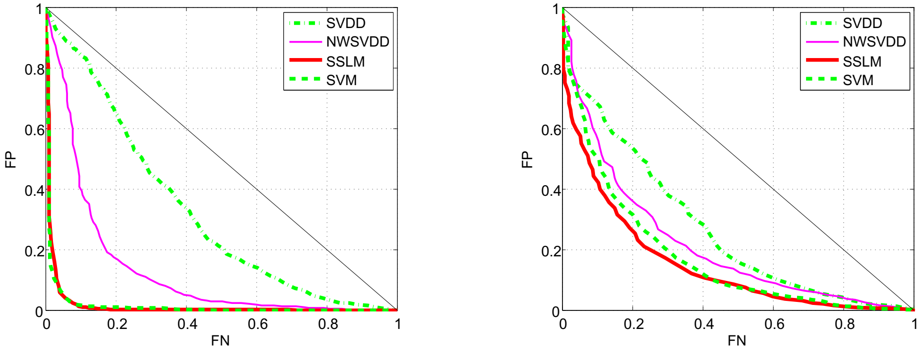
3 Drug Activity Prediction from Molecular Structure

The design of new medical drugs with desired chemical properties has a capital importance for the pharmaceutical industry. Several approaches are used in drug discovery, which can be grouped in three main categories: random screening of a large number of compound in a blind way, structural modifications of lead compounds and rational drug design [9]. Quantitative structure-activity (structure-property) relationships (QSAR/QSPR) constitute a methodology in the last category that is based on the fact that some properties of a set of molecules change with their molecular structure and therefore it is possible to find a relationship between this structure and the properties that the molecule exhibits. Once this relationship has been obtained it can be used to predict the properties of new, perhaps unknown, compounds.

Molecular descriptors used in QSAR can be empirical (derived from experimental data) or nonempirical. Among the nonempirical descriptors, the so-called topological indices have special relevance [10]. Topological indices are molecular descriptors derived from information on connectivity and composition of a molecule and can be easily derived from the hydrogen-suppressed molecular representation seen as a graph [11,12]. Some examples of topological indices are the popular Kier and Hall connectivity index [13] and Balaban index of average distance sum connectivity [14]. In this work, a set of 116 indices has been selected from three families considered that we will refer to as topological [15], the above mentioned Kier-Hall and the electro-topological or charge index[16]. Some experiments have been carried out using a reduced set formed by the 62 topological indices.

To properly assess the different predictors in this context, Receiver Operating Characteristic (ROC) curves and associate performance measures have been considered in this work [17]. Given a particular predictor whose output consists of a continuous value in a specified interval (as in this work), the ROC curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form (FP,TP). The performance of any classifier (with the corresponding threshold included) can be represented by a point in the ROC space. ROC curves move from the "all-inactive" point (0,0) which corresponds to the highest value of the threshold to the "all-active" point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A common method to compare classifiers is to calculate the area under



a) Antibacterial database

b) Analgesic database

Fig. 1. ROC curves corresponding to the different predictors considered

the ROC curve (AUC). The value of the AUC will always be between 0.5 and 1.0, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of 0.5. The AUC has some important statistical properties [17] and is frequently used as a global measure of predictiveness.

4 Experimental Results

Several comparative experiments have been carried out using a wide range of settings for the algorithms considered. Two specific datasets containing chemical compounds have been considered. First, an small dataset of 434 compounds using 62 topological indices and exhibiting (218) or not (216) antibacterial activity have been considered [2]. Also, a more challenging and realistic dataset with 973 compounds where 111 of them exhibit analgesic properties have been used. In this second database, all 116 descriptors have been used to represent the compounds [15]. More details about data and availability are given in previous referenced works. Moreover, the experimental protocol including coding of all algorithms closely follows these previous studies.

As the main goal consists of an empirical comparison, a relatively wide range of settings has been tried for all the algorithms considered. To obtain appropriately averaged performance measures the n-fold cross validation procedure with $n = 10$ has been repeated four times. As a performance measure for each fold, the full ROC curve has been computed along with its AUC measure. Both ROC curves and AUC measures have been averaged over the different blocks in the cross validation procedure [17] and are shown in Figure 1 and Table 2, respectively. Only the results corresponding to the best settings for each algorithm have been presented. These settings for each particular algorithm and database are specified in Table 1. For all algorithms, a Gaussian kernel has been used whose parameter has been fixed as $\sigma = 0.125$ according to several previous studies using the same databases [2].

By observing the ROC curves obtained with the best settings for the Antibacterial database in Figure 1 it can be seen that there is a very significant difference between the two best algorithms (SVM and SSLM) and the rest. It is not surprising that the SVDD algorithm gives the worst results because it does not use negative examples. On the contrary, the poor result corresponding to the NWSVDD method was relatively unexpected. When considering the Analgesic database the performance of all algorithms gets significantly lower in all cases. This is due both to the fact that the problem is considerably more difficult and also because the database is severely unbalanced. For this database, SVDD and NWSVDD methods give virtually the same results along the ROC curve and SVM gives only slightly better results. The SSLM method gives the best results except for a small range in the curve. The AUC values shown in Table 2 numerically characterize the differences of performance among the different methods over the two databases. The AUC value corresponding to the LDA method has also been included as a baseline.

The particular AUC values obtained for each database in each of the 4 times 10 cross validation steps have been put together and a nonparametric Friedman test followed by a post-hoc Holm test [18] has been performed. Table 3 shows the obtained average rankings and adjusted p-values when comparing each method to SSLM. According to this, it can be said that the SSLM gives the best AUC results at a significance level of $\alpha = 0.05$.

These results illustrate the fact that OC predictors with enough information (counterexamples) and flexibility (in particular using a margin to separate examples from counterexamples) are able to improve on good binary classifiers (SVM). Nevertheless, the amount of improvement attained is relatively moderate. Apart from this improvement on the overall performance, the one-class predictors are interesting also because of its ability to adapt to different situations. In particular, in specific applications as the ones considered in this paper, it is possible to adapt the predictors to specific operating ranges of the ROC curve that correspond to specific situations. In other words, instead of looking for a unique model that gives rise to a good ROC curve, we can learn a specific model that is good only in a small range in the curve. This capability of the

Table 1. Best parameters for each one of the algorithms on each database

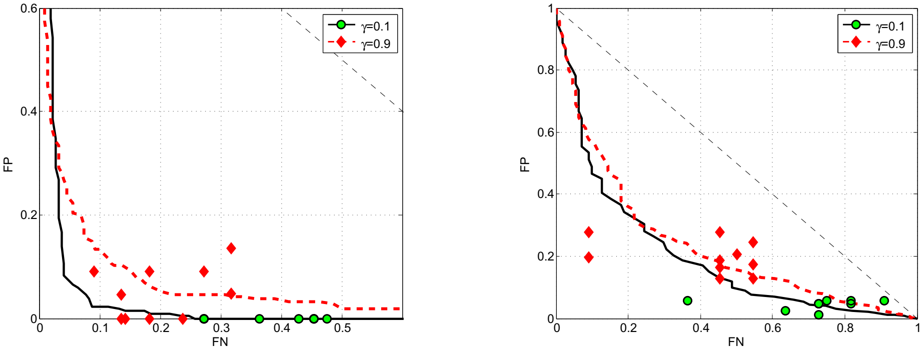
	SVM	SVDD	NWSVDD	SSLM
Antibacterial	$\nu = 0.0319$	$\nu = 0.25$	$\nu = 0.0125$, $\gamma = 0.25$	$\nu = 0.0001$, $\gamma = 0.1$, $\eta = 50$
Analgesic	$\nu = 0.1194$	$\nu = 0.9$	$\nu = 0.0040$, $\gamma = 0.35$	$\nu = 0.001$, $\gamma = 0.9$, $\eta = 40$

Table 2. AUC measure for each algorithm on each database

	LDA	SVM	SVDD	NWSVDD	SSLM
Antibacterial	0.966	0.976	0.686	0.871	**0.985**
Analgesic	0.834	0.829	0.732	0.788	**0.852**

Table 3. Average rankings and adjusted p-values

Algorithm	SSLM	SVM	LDA	NWSVDD	SVDD
Ranking	1.669	2.256	2.513	3.825	4.737
Adjusted p-value		0.03755	0.00221	$< 10^{-16}$	$< 10^{-32}$



a) Antibacterial database

b) Analgesic database

Fig. 2. Best one-class models (SSLM) obtained for each database by forcing the algorithm to minimize either FP or FN rates by forcing the parameter γ

models has not been fully exploited in this work but Figure 2 shows two specific models specialized at the different endings of the ROC curve. In these figures particular predictors obtained at each one of ten runs are shown along with the corresponding averaged curves. In the case of Antibacterial database, it is possible to obtain predictors able to minimize one of the two types of errors but at different rates. In the Analgesic database a similar behavior can be observed. In both cases, the variability in the false negative rate is higher than the one in false positive rate.

5 Concluding Remarks

In this work, several different one-class predictors have been applied to a particular challenging problem related to drug activity characterization. In particular, recently proposed one-class predictors using counterexamples and a separation margin have been shown to give very interesting solution for this kind of problems. The behavior of the different models has been characterized by their corresponding ROC curves and AUC measures. Apart from the overall performance results it has been shown that the models can be adapted to different specifications in terms of maximum rates of each type of error. Further work is currently directed towards the specific problem of obtaining one or several one-class predictors optimized at different specific error rates.

References

1. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining 2(2), 121–167 (1998)
2. Ferri, F.J., Diaz-Villanueva, W., Castro, M.: Experiments on automatic drug activity characterization using support vector classification. In: IASTED Intl. Conf. on Computational Intelligence (CI 2006), San Francisco, US, pp. 332–337 (2006)
3. Yepes, V., Pellicer, E., Ferri, F.: Profit forecasting using support vector regression for consulting engineering firms. In: 9th International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy (2009)
4. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
5. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)
6. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
7. Wu, M., Ye, J.: A small sphere and large margin approach for novelty detection using training data with outliers. IEEE Trans. Pattern Anal. Mach. Intell. 31(11), 2088–2092 (2009)
8. Cao, L.J., Lee, H.P., Chong, W.K.: Modified support vector novelty detector using training data with outliers. Pattern Recogn. Lett. 24(14), 2479–2487 (2003)
9. Gozalbes, R., Doucet, J., Derouin, F.: Application of topological descriptors in qsar and drug design: History and new trends. Current Drug Targets – Infectious Disorders 2, 93–102 (2002)
10. Katritzky, A.R., Gordeeva, E.V.: Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in qsar/sqpr research. J. Chem. Inf. Comput. Sci. 33, 835–857 (1993)
11. Basak, S., Bertelsen, S., Grunwald, G.: Application of graph theoretical parameters in quantifying molecular similarity and structure-activty studies. J. Chem. Inf. Comput. Sci. 34, 270–276 (1994)
12. Seybold, P., May, M., Bagal, U.: Molecular structure-propertiy relationships. J. Chem. Educ. 64, 575–581 (1987)
13. Kier, L.B., Hall, L.H.: Molecular Connectivity in Structure-Activity Analysis. John Willey and Sons, New York (1986)
14. Balaban, A.T.: Highly discriminating distance-based topological index. Chem. Phys. Lett. 89, 399–404 (1982)
15. Gálvez, J., García-Domenech, R., de Julián-Ortiz, J., Soler, R.: Topological approach to drug design. J. Chem. Inf. Comput. Sci. 35, 272–284 (1995)
16. Galvez, J., Garcia, R., Salabert, M., Soler, R.: Charge indexes. new topological descriptor. J. Chem. Inf. and Comp. Sciences 34, 502–525 (1994)
17. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)
18. García, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)