Bioorganic & Medicinal Chemistry Letters 25 (2015) 100–105

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Exploiting uncertainty measures in compounds activity prediction using support vector machines

Sabina Smusz [a,b], Wojciech Marian Czarnecki [c], Dawid Warszycki [a], Andrzej J. Bojarski [a,*]

[a] Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, Smetna 12, Kraków 31-343, Poland
[b] Faculty of Chemistry, Jagiellonian University, R. Ingardena 3, Kraków 30-060, Poland
[c] Faculty of Mathematics and Computer Science, Jagiellonian University, S. Lojasiewicza 6, Kraków 30-348, Poland

ARTICLE INFO

Article history:
Received 19 June 2014
Revised 31 October 2014
Accepted 1 November 2014
Available online 7 November 2014

Keywords:
Uncertainty of in vitro tests
Machine learning
Support Vector Machine
Weighting protocol
ChEMBLdb

ABSTRACT

The great majority of molecular modeling tasks require the construction of a model that is then used to evaluate new compounds. Although various types of these models exist, at some stage, they all use knowledge about the activity of a given group of compounds, and the performance of the models is dependent on the quality of these data. Biological experiments verifying the activity of chemical compounds are often not reproducible; hence, databases containing these results often possess various activity records for a given molecule. In this study, we developed a method that incorporates the uncertainty of biological tests in machine-learning-based experiments using the Support Vector Machine as a classification model. We show that the developed methodology improves the classification effectiveness in the tested conditions.

Molecular modeling methods (although abstract) always make use of experimental data. They either constitute basis upon which the model is constructed (e.g., pharmacophore models)[1] or they are used as a training/verification element (e.g., docking).[2] At some stage, they all use knowledge about the activity of a given group of compounds.

There are a number of databases that provide quantitative information on the biological activity of chemical compounds, such as ChEMBL,[3] PDSP,[4] and PubChem,[5] among others. However, due to the inconsistency of the results obtained from in vitro experiments, for some compounds, there is more than one provided K_i (or equivalent parameter) value. For example, in the case of cocaine, there are 815 different activity records (with differences also occurring within the same assay conditions—for example, 22 activities reported for the cocaine potency towards D_2 receptor) in the ChEMBL database.

In this study, we developed a modification of the Support Vector Machine (SVM)[6] that takes the uncertainty of biological experiments into account. This method was verified using data from the ChEMBL database on 25 protein targets and PaDEL fingerprints[7] to represent the compound structures. The approach was compared with standard experiments that did not consider the uncertainty of the in vitro data, and its superiority over standard methods was proven.

We propose to use knowledge on the uncertainty of the activity of compounds by exploiting the variances in their K_i values that have been reported in activity databases, and we will show how this information can be incorporated into the Support Vector Machine's optimization problem. Thus, one can use nearly any existing implementation of the SVM to perform such analysis. Let us denote our samples from the dataset as triplets (x_i, y_i, a_i), where x_i is a compound's fingerprint representation, y_i is its activity class (+1: active or −1: inactive), and a_i is the set of K_i values obtained during experimental activity testing. We can reformulate the Support Vector Machine's optimization problem to the following, which exploits knowledge of the activity uncertainty:

$$\underset{w,b}{\text{minimize}}\ \tfrac{1}{2}||w||^2 + C\sum_i^N \xi_i$$

$$\text{subject to } y_i(\langle w, x_i\rangle - b) \geqslant 1 - \xi_i - \text{var}(a_i)\xi_i, \quad i = 1, \ldots, N$$

where ξ_i are slack variables that measure how far the ith point is from the correct classification and N is the number of molecules used for predictive model construction (compounds present in the training set).

As a result, compounds with high K_i variance are less important during model construction (they 'are allowed to' violate the separating margin by the distance proportional to their values' variances). From a practical point of view, this type of optimization procedure can be solved using samples-weighted SVM, with the i'th sample weight equal to $(1 + \text{var}(a_i))^{-1}$. It is worth noting that,

* Corresponding author.

http://dx.doi.org/10.1016/j.bmcl.2014.11.005
0960-894X/

although it is a modification of the optimization problem, this method does not require any custom implementation, and in fact, any existing SVM library that supports this (samples) weighting (including SVMlight,[8] scikit-learn,[9] etc.) can be used. Figure 1 shows how introduction of knowledge about uncertain data helps the SVM to find the best separating hyperplane. It is easy to notice that for the entire certain dataset (where each compound has only one annotated K_i value), this problem degenerates to a simple SVM optimization.

The applicability of the method was verified using 25 protein targets (serotonin receptors 5-HT_{2A},[10] 5-HT_{2C},[11] 5-HT_6,[12] and 5-HT_7,[13] cyclin dependent kinase 2 (CDK2),[14] muscarinic receptor M_1,[15] MAP kinase ERK2,[16] acetylcholinesterase (AChE),[17] adenosine receptor A1,[18] alpha-2A adrenergic receptor (α2AR),[19] atrial natriuretic peptide receptor (ANP),[20] beta-1 adrenergic receptor (beta1AR),[21] beta-3 adrenergic receptor (beta3AR),[21] cannabinoid CB1 receptor,[22] delta opioid receptor (DOR),[23] dopamine receptor D_4,[24] histamine receptor H_1,[25] histamine receptor H_3,[26] HIV integrase (HIVi),[27] insulin receptor (IR),[28] tyrosine kinase ABL,[29] human leukocyte elastase (HLE),[30] norepinephrine transporter (NET),[31] phosphodiesterase 4A (PDE4A),[32] and vasopressin 1A receptor (V1a)[33]) which were represented by the following fingerprints generated with the use of the PaDEL-Descriptor: E-state Fingerprint (EstateFP, 79 bits),[34] Extended Fingerprint (ExtFP, 1024 bits),[35] Klekota–Roth Fingerprint (KlekFP, 4860 bits),[36] MACCS Fingerprints (MACCSFP, 166 bits),[37] Pubchem Fingerprint (PubchemFP, 881 bits),[38] and Substructure Fingerprint (SubFP, 308 bits), respectively.[39]

Compounds with experimentally verified activity towards the selected proteins were obtained from the ChEMBL database. Only molecules whose activities were quantified in K_i or IC_{50} (after careful data analysis and examination of the most common protein concentrations, it was assumed that $K_i = IC_{50}/2$) and were tested in assays on human, rat-cloned or native receptors were taken into account. The compounds were considered active when the median value of all K_i values provided for a particular instance was lower than 100 nM, and they were considered inactive when the median value was greater than 1000 nM. The number of compounds from each group for the selected targets together with an analysis of the reliability of the biological data is shown in Table 1 (detailed analysis of K_i variance is included in the Supplementary materials section). The table shows that there are great differences in terms of reliability of the biological data between the various targets (on average, nonzero variances occurred for 5-HT_{2C} in less than 5% of cases vs HIVi, where different K_i values were found for almost 50% of cases) and the various groups of compounds—for example, for the activity threshold applied in this study, over 10% of active 5-HT_7 ligands had more than one provided K_i value, whereas only approximately 2% of inactive compounds had uncertain K_i values (in terms of the number of different provided values).

Table 1
The number of active and inactive compounds for each target used in the study with an analysis of the reliability of the biological data.

Protein	Actives	Inactives
5-HT_{2A}	1836 (6.3%)	852 (2.6%)
5-HT_{2C}	1211 (5.3%)	927 (2.6%)
5-HT_6	1491 (8.6%)	342 (0.5%)
5-HT_7	705 (10.6%)	340 (1.8%)
CDK2	741 (4.2%)	1462 (2.1%)
M_1	760 (19.7%)	939 (11%)
ERK2	72 (1.4%)	958 (1.9%)
AChE	1147 (11.9%)	1804 (4.8%)
A1	1789 (8.2%)	2286 (3.9%)
α2AR	364 (8.0%)	283 (2.5%)
ANP	114 (0.2%)	142 (1.1%)
Beta1AR	195 (1.8%)	477 (0.1%)
Beta3AR	111 (0.9%)	133 (0.0%)
CB1	1964 (15.3)	1714 (4.6%)
DOR	2535 (9.1%)	1992 (1.9%)
D_4	1034 (11.1%)	449 (1.8%)
H_1	636 (8.5%)	546 (1.3%)
H_3	2706 (7.4%)	313 (1.7%)
HIVi	102 (32.4%)	915 (53.6%)
IR	147 (3.4%)	1139 (1.7%)
ABL	409 (7.9%)	582 (3.2%)
HLE	820 (4.0%)	610 (1.4%)
NET	1738 (15.3%)	1299 (5.3%)
PDE4A	303 (11.0%)	82 (10.9%)
V1a	467 (12.3%)	300 (1.7%)

Percentages in parentheses denote the number of compounds with more than one provided K_i value.

Due to high-class imbalance (differences between the number of active and inactive compounds within each dataset), the classification effectiveness was evaluated using a balanced quality measure, that is, balanced accuracy (BAC):

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$$

To maximize the score under some balanced measure (such as BAC or Matthew's Correlation Coefficient),[40] the class-based weighting scheme in the SVM formulation should be used. By default, SVM maximizes the accuracy measure (the percentage of correct predictions), which is not balanced and favors the larger class. The class-weighting technique is a well-known approach based on weighting samples by the number inversely proportional

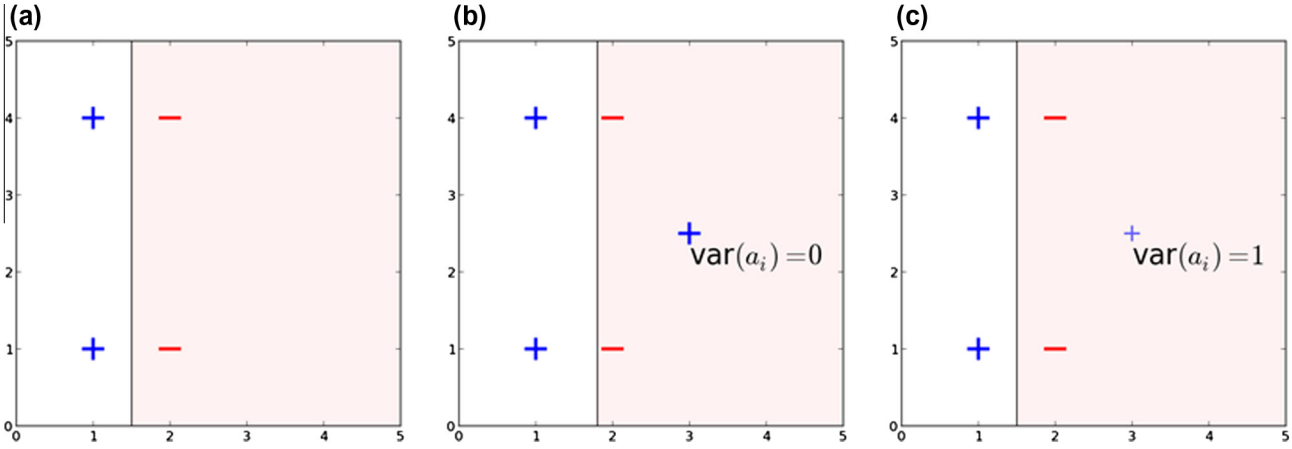

Figure 1. Visualization of the activity-uncertainty-weighted SVM with: (a) linearly separable data, (b) linearly nonseparable data and variance of the rightmost point a_i equal to 0, and (c) linearly nonseparable data and variance of the rightmost point a_i equal to 1; the higher the variance, the less important a particular sample.