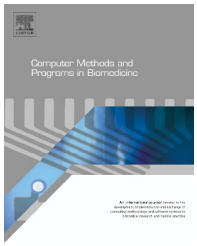
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 51–60

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Drug/nondrug classification using Support Vector Machines with various feature selection strategies

*Selcuk Korkmaz**, *Gokmen Zararsiz*, *Dincer Goksuluk*

Hacettepe University, Faculty of Medicine, Department of Biostatistics, 06100 Sihhiye, Ankara, Turkey

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form
15 August 2014
Accepted 27 August 2014

Keywords:
Support Vector Machines
Molecular descriptors
Feature selection
Drug discovery
Machine learning

ABSTRACT

In conjunction with the advance in computer technology, virtual screening of small molecules has been started to use in drug discovery. Since there are thousands of compounds in early-phase of drug discovery, a fast classification method, which can distinguish between active and inactive molecules, can be used for screening large compound collections. In this study, we used Support Vector Machines (SVM) for this type of classification task. SVM is a powerful classification tool that is becoming increasingly popular in various machine-learning applications. The data sets consist of 631 compounds for training set and 216 compounds for a separate test set. In data pre-processing step, the Pearson's correlation coefficient used as a filter to eliminate redundant features. After application of the correlation filter, a single SVM has been applied to this reduced data set. Moreover, we have investigated the performance of SVM with different feature selection strategies, including SVM–Recursive Feature Elimination, Wrapper Method and Subset Selection. All feature selection methods generally represent better performance than a single SVM while Subset Selection outperforms other feature selection methods. We have tested SVM as a classification tool in a real-life drug discovery problem and our results revealed that it could be a useful method for classification task in early-phase of drug discovery.

1. Introduction

Traditionally, drug discovery process starts with the identification of the disease-associated protein [1] and then this process is followed by testing of the disease protein against thousands of known and new compounds to find lead compounds that can interact with the target protein and show the potential effectiveness against disease. These lead compounds can serve as candidates for the drug to be further analyzed in pre-clinical studies. Since thousands of compounds screened from compound libraries in this step, virtual screening (VS) can be used to search libraries in order to identify structures, which are most likely to bind to a target protein [2,3]. Since VS is a computational filter, which reduces the size of a chemical library to be screened experimentally, it can reduce time and effort in finding lead compounds and thereby saves money. Early-phase VS often employs to eliminate potentially unwanted molecules (i.e. inactive or toxic) from a compound library [4]. Therefore, machine-learning (ML) methods can be used for VS by analyzing the structural features of molecules of known activity or inactivity [5].

The main issue in early-phase of drug discovery process is the evaluation of drug compounds. Hereby, drug compounds have been studied from different perspectives [6] including prediction of oral bioavailability [7–10], drug-like [11–18] and

* *Corresponding author*. Tel.: +90 312 305 1467; fax: +90 312 305 1459.
E-mail addresses: selcuk.korkmaz@hacettepe.edu.tr, selcukorkmaz@hotmail.com (S. Korkmaz).

http://dx.doi.org/10.1016/j.cmpb.2014.08.009
0169-2607/

lead-like [19] compounds, number [20] and topology of rings [21], molecular frameworks [22,23] and fragments [14,24–28]. Over a decade, various ML methods have been applied to biology, chemistry and drug discovery [29]. Supervised ML methods such as linear discriminant analysis [30] and decision trees [31] were used to predict structural properties of compounds. Furthermore, logistic regression [6], Bayesian networks [11] or artificial neural networks [12] have been used to distinguish between drugs and nondrugs. In addition to the activity studies, principal component analysis, Bayesian networks, neural networks and Support Vector Machines (SVM) were used in various chemogenomic studies [6].

SVM is one of the most widely used ML methods. Recently, it is used in a variety of drug discovery applications. There are some studies tried to design new kernel functions for SVM to combine compound structures with other data [32–38] and transfer similarity calculations into a combined feature space [39–43]. In addition to kernel design studies, SVM was used to predict compounds with single-target activity [4,40,44–58] and multi-target activity [59–63], and different compounds properties [64–76]. Furthermore, SVM methodology was used to predict drug-likeness score for targets [77], target–ligand interactions [78] and protein–ligand binding affinities [79–81]. More detailed information can be found in Heikamp and Bajorath [29].

In the present study, we applied SVM to a real-life drug discovery problem, particularly, the comparison of active against inactive molecules for screening large compound libraries. We built our SVM models with incorporation of various feature selection methods, including SVM/Recursive Feature Elimination (SVM/RFE), Wrapper Method (WM) and Subset Selection (SS). Additionally, to check for external validity of the study, we compared our results with literature in the field.

2. Methodology

2.1. Data sets

The training and test sets of compounds have been taken from a different publication [6]. Based on that study, the training set contained 311 *drugs* and 320 *nondrugs* and the test set, which was a independent set of compounds, contained 98 *drugs* and 118 *nondrugs*. The data matrix consisted of 34 molecular descriptors as follows: log*P* (the logarithm of the octanol/water partition coefficient), *NHA* (number of heavy atoms), *MW* (molecular weight), *NoC* (number of carbons), *AC* (atom count), *HC* (hydrogen count), *HBDC* (hydrogen bond donor count), *HBAC* (hydrogen bond acceptor count), *RBC* (rotatable bond count), *AlRC* (aliphatic ring count), *ArRC* (aromatic ring count), *AAC* (aromatic atom count), *BC* (bond count), *RC* (ring count), *MSA* (molecular surface area), *PSA* (polar surface area), *APSA* (apolar surface area), *MP* (molecular polarizability), *WI* (Wiener index), *BI* (Balaban index), *HI* (Harary index), *hWI* (hyper-Wiener index), *PI* (Platt index), *RI* (Randic index), *SI* (Szeged index), *WPI* (Wiener polarity index) and 8 more ligand efficiency indices; ΔG_{Bind}/*NHA*, ΔG_{Bind}/*MW*, ΔG_{Bind}/*NoC*, ΔG_{Bind}/*PSA*, ΔG_{Bind}/*MSA*, ΔG_{Bind}/*APSA*, ΔG_{Bind}/*WI*, ΔG_{Bind}/*P*, where ΔG_{Bind} is the binding energy and *P* is the octanol/water partition coefficient. More detailed information about the molecular descriptors and the data sets can be found in Garcia-Sosa et al. [6].

2.2. Support Vector Machines

SVM is a popular classification tool, which originally presented by Vapnik and his co-workers and has taken great interest from science community because of its strong mathematical background and excellent empirical successes. SVM is also capable of nonlinear classification and handling high-dimensional data, thus applied in many fields such as computational biology, text classification, image segmentation and cancer classification [82,83].

In binary classification problems, let $\{x_1, \ldots, x_n\}$ is a given training data that are n-dimensional vectors in some space ($x_i \in \mathbb{R}^n$) and $\{y_1, \ldots, y_n\}$ are their class labels where $y_i \in \{-1, +1\}$. The aim here is to find a hyperplane and obtain an equation, which separates the training data into two parts that all data points with same class labels exist on the same side of the hyperplane. Here, the data points that are closest to the hyperplane in both side is called support vectors and the objective of SVM is to maximize the margin $1/w$, which is the distance between two support vectors or minimize $w^2/2$ equivalently. SVM takes advantage of Lagrange multipliers and Karuck Kuhn Tucker conditions to overcome this optimization problem.

When the data is linearly non-separable, slack variables $\{\xi_1, \ldots, \xi_n\}$, which is a penalty introduced by Cortes and Vapnik [84], can be used to allow misclassified data points, where $\xi_i > 0$. Moreover, in many classification problems, the separation surface is nonlinear. SVM deals with this problem by mapping the input vectors to a high-dimensional space by using kernel functions (e.g. polynomial, radial-based, sigmoidal). A detailed description of SVM algorithm can be found in [82].

2.3. Data pre-process and feature selection

Since data pre-process and feature selection (FS) improve the prediction performance of predictors, provide faster and more cost-effective predictors and offer a better understanding of the underlying process that generated the data, they are the most crucial steps in ML methods. In pre-processing step, we applied the following two steps to our training set: (i) we have split the 34 molecular descriptors into 6 categories (8 topological indices, 10 atom and bond counts, 3 size and shape descriptors, 2 pharmacophore descriptors, 3 physical descriptors and 8 ligand efficiency descriptors) based on their properties [85] (see Fig. 1), (ii) we have computed the Pearson's correlation coefficient for the entire pairs of descriptors in each category. For all cross terms that were either strongly positive or negative ($r > 0.90$ or $r < -0.90$) correlated with each other, we have discarded one descriptor based on *t*-test. Thus, we have selected the descriptor that provides more information for classification. In addition to the correlation filter, we used three FS methods in this study, including SVM–Recursive Feature Elimination (SVM–RFE), Wrapper Method (WM) and Subset Selection (SS).