Pattern recognition approach to classifying CYP 2C19 isoform

Bartosz Krawczyk*

Wrocław University of Technology, Department of Systems and Computer Networks, Wybrzeże Wyspianskiego 27, 50-370 Wrocław, Poland.

Received 09 September 2011; Accepted 27 October 2011

Abstract: In this paper a pattern recognition approach to classifying quantitative structure-property relationships (QSPR) of the CYP2C19 isoform is presented. QSPR is a correlative computer modelling of the properties of chemical molecules and is widely used in cheminformatics and the pharmaceutical industry. Predicting whether or not a particular chemical will be metabolized by 2C19 is of primary importance to the pharmaceutical industry. This task poses certain challenges. First of all analyzed data are characterized by a significant biological noise. Additionally the training set is unbalanced, with objects from negative class outnumbering the positives four times. Presented solution deals with those problems, additionally incorporating a throughout feature selection for improving the stability of received results. A strong emphasis is put on the outlier detection and proper model validation to achieve the best predictive power.

Keywords: Pattern recognition • Machine learning • Medical informatics • Chemoinformatics• Unbalanced training set • Feature selection

© Versita Sp. z o.o

1. Introduction

The pharmaceutical industry exists for, and profits from, the discovery and development of new medicines. Currently the costs of introducing a new drug to the medical market are huge, reported to be reaching 1.7 billion US dollars and approximately about 10-15 years of constant clinical research (information supplied by the Simulation Plus, Inc. [1]). Up until the late 90’s, the pharmaceutical industry was mainly using slow and costly experimental work in laboratories and clinics. A desire to eliminate many unnecessary costs has accelerated the emergence and acceptance of the science of cheminformatics, roughly defined as data mining in chemical space [2]. A concept that similar chemical compounds display similar properties, is the basis to the idea of taking existing solubility data and building new statistical correlative models. This allows to create a mapping between the structure of the chemical molecules and their water solubility. This computer modelling of properties of the chemical molecules is known as Quantitative Structure-Property Relationships/Quantitative Structure-Activity Relationships and abbreviated QSPR/QSAR. These models are used for fast screening out the potential drug candidates with wanted or unwanted properties. Therefore they are of a major importance to the pharmaceutical companies or to medical and clinical researchers.

A perfect situation will occur when we would have a database with sufficient number of QSAR/QSPR models. Then a pharmaceutical researcher would not have to even synthesize new compounds whose properties were predicted as unpromising. Of course, this situation is purely theoretical and never happens. All QSAR/QSPR models are far from ideal and carry with them a relatively significant prediction error. That leaves a wide area for introducing new and efficient statistical methods to deal with such problems.

In this paper a pattern recognition approach to classifying chemical molecules into substrates and non-substrates of the CYP 2C19 isoform of the cytochrome P450 enzyme in human is shown. The aim is

* E-mail: bartosz.krawczyk@pwr.wroc.pl
to predicting whether or not a particular chemical will be metabolized by 2C19, as well as other major isoforms of CYP P450. This knowledge is of primary importance to the pharmaceutical industry. This task poses certain difficulties. Analyzed data are characterized by a significant biological noise and the training set is unbalanced. This paper presents an approach that efficiently solves those problems.

The data for this task was supplied by Simulation Plus, Inc. [1] as a part of open pattern recognition and data mining challenge.

The content of the work is as follows. In the next subsection related works in this field are discussed. Then a pattern recognition task, Support Vector Machine (SVM) algorithm and chemoinformatics background are presented shortly. Next general idea and detailed information about proposed method are introduced. The results of computer experiments are described in section 5. The last section concludes the paper.

1.1. Related works

Problem of the unbalanced data sets was addressed in a number of publications. Apart from presented in the further sections SMOTE approach [3] various alternation of this algorithm were introduced in recent years [4,5]. But in case when dealing with non-Euclidean and nonmetric data, the application of those methods are impossible. For such data, e.g. time series, a ghost point approach was developed [6]. A different, yet interesting approach of boosting SVM was presented in [7]. Neural Networks were also adapted to this problem with two most representative approaches. First one considered implementing new, dynamic threshold learning algorithm [8]. Second concentrated on transforming an unbalanced problem into a set of symmetrical, two-class problems, each of which can be solved by a simple neural network [9].

In chemoinformatics a computer-assisted synthesis design (CASD) was seen as a highly interesting challenge and as a field for applying artificial intelligence techniques [10]. In 1980 a very important work for this filed, by Lindsay et al. was introduced [11]. Since then there is an active group dedicated to this area [12]. One of the first work, which introduced the interdisciplin ary application of machine learning techniques to the structure-activity modelling was [13]. Few years later a first summary paper, describing state-of-the-art in this field was printed [14]. It can be clearly seen that then only the feature selection was considered as an important pre-processing step, which was further discussed in [15]. In [16] authors stressed that using different pre-processing can further improve the overall quality of structure-activity classification.

2. Materials and methods

2.1. Pattern recognition

The aim of the pattern recognition is to classify a given object to the one of the predefined categories, on the basis of observation of the features describing it [17]. The object and its attributes are presented as a feature vector $x \in X$.

The pattern recognition algorithm maps the feature space $X$ to the set of class labels $M$.

$$\psi : X \rightarrow M$$

The mapping (1) is established on the basis of examples included in a learning set or rules given by experts. The learning set consists of learning examples, i.e. observation of features described object and its correct classification. The idea of classification systems is depicted in Figure 1.

2.2. Support Vector Machine Classifier

Support vector machine (SVM) is a supervised learning method that recognize patterns in the analyzed data, used for classification and regression [18]. The standard SVM is a non-probabilistic binary linear classifier.

The idea of SVM can be explained shortly as follows. A support vector machine constructs a set of hyperplanes in a high dimensional feature space, which can be used for data separation, and thus for classification. A hyperplane achieves a good separation that if it has the largest distance to the nearest training data points of any class (so-called functional margin). In general the larger the margin means the lower generalization error of the classifier.

It often happens that in real-life scenarios the sets are very complex and it is impossible to separate them linearly. To deal with this inconvenience it was proposed that the original finite dimensional space be mapped into a much higher dimensional space, which will make the linear separation possible. SVM use a fast mapping into a larger space so that cross products may be computed efficiently for the variables in the original space.