Journal of Chromatography B, 910 (2012) 84–94

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography B

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

QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions☆

Mohammad Goodarzi[a], Richard Jensen[b], Yvan Vander Heyden[a,*]

[a] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium
[b] Department of Computer Science, Aberystwyth University, Aberystwyth, Wales, UK

ARTICLE INFO

Article history:
Received 7 September 2011
Accepted 17 January 2012
Available online 24 January 2012

Keywords:
QSRR
Chromatographic retention
ACO
MLR
SVM
Relief method

ABSTRACT

A Quantitative Structure-Retention Relationship (QSRR) is proposed to estimate the chromatographic retention of 83 diverse drugs on a Unisphere poly butadiene (PBD) column, using isocratic elutions at pH 11.7. Previous work has generated QSRR models for them using Classification And Regression Trees (CART). In this work, Ant Colony Optimization is used as a feature selection method to find the best molecular descriptors from a large pool. In addition, several other selection methods have been applied, such as Genetic Algorithms, Stepwise Regression and the Relief method, not only to evaluate Ant Colony Optimization as a feature selection method but also to investigate its ability to find the important descriptors in QSRR. Multiple Linear Regression (MLR) and Support Vector Machines (SVMs) were applied as linear and nonlinear regression methods, respectively, giving excellent correlation between the experimental, i.e. extrapolated to a mobile phase consisting of pure water, and predicted logarithms of the retention factors of the drugs ($\log k_w$). The overall best model was the SVM one built using descriptors selected by ACO.

1. Introduction

For many years, the separation of drugs has been a critical and important stage in analytical chemistry and pharmaceutical science. One of the most applied techniques is High-Performance Liquid Chromatography (HPLC), which is able to analyze a wide polarity range of acidic, basic and neutral compounds. High-Performance Liquid Chromatography is well recognized as a powerful, fast, selective and highly efficient technique, successfully employed for the separation and determination of many drugs [1]. To perform separations, a broad range of chromatographic stationary phases provide meaningfully different retention and selectivity. However, the mechanisms of retention are not always entirely known [2,3]. The choice of the stationary phase is very important and is based on user knowledge or on chromatographic tests to select columns with similar or dissimilar characteristics (selectivities).

The prediction of the physicochemical behavior of compounds, such as chromatographic retention, is useful for estimating, for instance, how well two similar substances will be distinguished in a given separation system, at the moment standards are not (yet) available in the drug development process. Quantitative Structure-Retention Relationship modeling has been utilized for the prediction of retention and migration behaviors [4–9]. In the resulting models a retention parameters is modeled as a function of molecular descriptors. It should be noted that QSRR is a kind of Quantitative Structure-Property Relationship (QSPR) study. Put et al. [9] have performed Classification And Regression Tree (CART) analysis as a QSRR study of the chromatographic retention of 83 drugs. CART selected three descriptors: a hydrophobicity parameter ($\log P$), the hydrophilic factor (Hy) [10] and the total path count (TPC) [11] from 266 calculated descriptors, to predict chromatographic retention. CART divided the retentions of the 83 molecules into five classes called very low, low, intermediate, high and very high retention [9].

In the present study, we have performed regression instead of classification. One of the most important stages, not only in classification but also in regression, is feature selection. As many pattern recognition and regression techniques were originally not designed to cope with large amounts of irrelevant features (e.g. given molecular descriptors), combining them with feature selection techniques has become a necessity in many applications [12–14]. The application of feature selection methods has several goals: firstly, to avoid overfitting and improve model performance;

☆ This paper belongs to the Special Issue Chemometrics in Chromatography, Edited by Pedro Araujo and Bjørn Grung.

* Corresponding author. Tel.: +32 24774723; fax: +32 24774735.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

1570-0232/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2012.01.012

secondly, to provide faster and more cost-effective models; and thirdly, to acquire a deeper insight into the underlying processes that generated the data, and to identify important variables that have an intuitive physical interpretation [15]. In this study we mainly focus on the first two goals and less on the latter.

Recently, Swarm Intelligence has been used in different fields of study for the purpose of feature selection [16]. One interesting method is Ant Colony Optimization (ACO). ACO [16–18] is based on the behavior of real ants that are capable of finding the shortest route between a food source and their nest by means of pheromone deposition, without the use of visual information and hence possessing no global world model, while being able to adapt to changes in the environment. If a sudden environmental change occurs (e.g. a large obstacle appears on the shortest path), the ants can respond to this and will eventually converge to a new path. Based on this idea, artificial ants can be deployed to solve complex optimization problems via the use of artificial pheromone deposition. ACO is particularly attractive for feature selection as there seems to be no heuristic that can guide incremental search to the optimal subset of features. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space. The ACO-based Fuzzy-Rough Set feature selection method has been applied recently for the first time in QSAR [19], giving excellent results for a class of glycogen synthase kinase-3β inhibitors.

Another important item which affects the prediction ability of any QSRR model is the choice of the regression technique for correlating descriptors with the experimental chromatographic retention. The significance of simple Multiple Linear Regression (MLR) in QSAR and QSRR has received attention from the literature [20,21], while accounting for non-linearity in the building of QSAR and QSRR models has also played an important role in the accuracy of activity and retention predictions, respectively [22,23]. It should be noted that in this study Support Vector Machines were used as nonlinear modeling technique. However, for the evaluation of ACO and to assess the ability of other feature selection methods, we have also used Genetic Algorithms (GAs), the Relief method and Stepwise Regression to select relevant variables in the construction of different QSRR models.

2. Theory

2.1. Feature selection and Ant Colony Optimization

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems feature selection is a must because of the abundance of noisy, irrelevant or misleading features. The usefulness of a feature or feature subset is determined by both its relevancy and its redundancy. A feature is said to be relevant if it is predictive for the decision feature(s) (i.e. dependent variable(s) here retention expressed as $\log k_w$), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features (for instance, different $\log P$ estimates may be highly correlated). Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other. However, the complexity of locating such a globally optimal subset of features is usually prohibitive, which motivates the use of more advanced search techniques, such as ACO.

For ACO-based feature selection, the process begins with the generation of a number of ants, placed randomly on a graph that represents every possible combination of features. Here, each node corresponds to a dataset feature and each edge permits the traversal of an ant from one feature to another (Fig. 1). An amount of virtual pheromone (a real number in ([0,1]) is associated with each edge that indicates the popularity of this particular traversal by past ants. Ants then traverse the graph, making probabilistic decisions as to which nodes to visit based on this virtual pheromone and also a heuristic desirability measure, until a traversal stopping criterion is satisfied. This is typically when the heuristic measure has reached a pre-calculated global optimum for the data. If the criterion is not satisfied, the virtual pheromone on the edges is updated based on ant traversals, a new set of ants is created and the process iterates once more. More details and definitions can be found in [17].

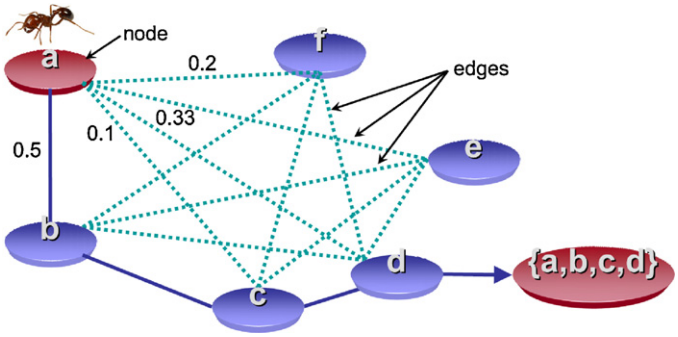

Fig. 1. ACO representation of feature selection. Nodes *a* to *f* represent features. The path highlighted (in blue) indicates the path taken by one ant and the resulting feature subset. The numbers on the edges are example of virtual pheromone. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2.2. The Relief method

In the Relief method [24,25], each feature is given a relevance weighting that reflects its ability to discern between decision class labels. It thus first was applied on classification problems. A user-specified threshold determines the number of sampled objects used for constructing the weights. For each sampling, an object **x** is randomly chosen, and its nearest neighbor of the same class and nearest neighbor of a different class are calculated. Based on these neighbors, the feature weights are updated such that more weight is given to features that discriminate the object from neighbors of different classes. The user must supply a threshold which determines the level of relevance that feature weights must surpass in order to be finally chosen. The method has been extended to enable it to handle inconsistency, noisy and multi-class datasets [26]. Relief has also been extended to handle continuous decision variables (e.g. retention parameters). Instead of requiring the exact knowledge whether two objects belong to the same class or not, which is not applicable in regression problems, the relative distance between the predicted values of two objects (compounds) is used in order to calculate feature weightings.

2.3. Support Vector Machine regression

Support Vector Machine (SVM) is a new and very promising classification and regression technique developed by Vapnik [27]. Here, we give only a brief introduction to its main principle. Given a training data set of compounds $\{(\mathbf{x}_1, y_1, \ldots, (\mathbf{x}_l, y_l)\}_{i=1}^{l}$, where $\mathbf{x}_i \in \mathbf{X} \subseteq R$ is the *i*th input data point in input space (a descriptor) and $y_i \in \mathbf{y} \subseteq R$ is the associated output value of $\mathbf{x}_i$ (retention parameters). Initially SVM considered classification problems of two classes. An SVM model is a representation of the samples as points in space, mapped in such a way that the two classes are separated by a gap that is as wide as possible. Because the classes are not always linearly separable in the initial data space (of the descriptors), the technique