Chemometrics and Intelligent Laboratory Systems 149 (2015) 10–16

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Efficient variable selection batch pruning algorithm for artificial neural networks

Vasyl Kovalishyn [a,*], Gennady Poda [b,c,**]

[a] Department of Medical and Biological Research, Institute of Bioorganic Chemistry & Petrochemistry, 1 Murmanska Street, Kyiv 02660, Ukraine
[b] Drug Discovery Program, Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, ON M5G 0A3, Canada
[c] Leslie Dan Faculty of Pharmacy, University of Toronto, 144 College Street, Toronto, ON M5S 3M2, Canada

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form 12 October 2015
Accepted 13 October 2015
Available online 22 October 2015

Keywords:
Artificial neural networks (ANN)
Associative neural network (ASNN)
Batch pruning algorithm (BPA)
Chemometrics
k-Nearest neighbors (*k*-NN)
Self-organizing map (SOM) of Kohonen
Machine learning
Variable selection

ABSTRACT

Here we report a novel, fast and efficient algorithm for variable selection, the batch pruning algorithm (BPA). The method combines the artificial neural networks (ANN) ensemble learning and self-organized map (SOM) of Kohonen for clustering of descriptors, followed up with a selection of an optimal smaller subset of descriptors from each cluster based on calculated sensitivity of input neurons. BPA was validated on two publicly available, structurally diverse datasets: 584 inhibitors of *M. Tuberculosis* (MTB) growth and 1015 phosphodiesterase type 4 (PDE4) inhibitors. BPA was able to identify a smaller subset of 5% of molecular descriptors (out of about 1200 calculated with Talete Dragon) 50–100 times faster compared to conventional stepwise pruning methods (SPM), and yielded QSAR models of similar or slightly better accuracy as measured by Q^2 (0.73–0.77), RMSE (0.50–0.72) and MAE (0.36–0.57). 97% of compounds were predicted within 1 log unit. It took only 1.47 h to find the best set of descriptors by BPA compared to 119 h by ANN SPM for the MTB dataset, and 3.0 h compared to 237 h for the PDE4 set. Due to its high predictive accuracy and speed, BPA may find wide applicability in building better machine learning models to predict activity, selectivity, physical and ADMET properties for large datasets, and a large number of descriptors within reasonable time.

1. Introduction

Machine learning techniques are widely used in chemometrics and modern drug discovery to analyze spectra, data modeling and for prediction of activity, selectivity, physical, absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of chemical entities. The most popular methods in chemometrics are principal component analysis (PCA), multiple linear regressions (MLR), partial least squares (PLS), *k*-nearest neighbors (*k*-NN), random forest (RF), recursive partitioning (RP), support vector machines (SVM), and artificial neural networks (ANN). To build a predictive model one has to calculate a large number of molecular descriptors, varying from simple atomic counts to complicated molecular properties. Hopefully, some of the calculated descriptors will provide a predictive (ideally, causative; not necessarily linear) correlation or a trend with activity or other important characteristics [1]. Nowadays, it is possible to quickly generate thousands of molecular descriptors and fingerprints. However, a lot of them do not provide any significant correlation (or trend) with the activity, thus carrying the "noise", or are highly correlated with each other and carry similar information with some variability [2]. Inappropriate variable selection from a large pool of descriptors can result in "chance correlations" (non-predictive) and yield models with little or no predictive ability. Several statistical methods have been used to select the appropriate set of molecular descriptors for model building. Traditionally, methods such as principal component analysis (PCA) or partial least squares (PLS) were employed [3]. More recently, due to increased computer power, artificial intelligence methods have become more popular [4–7]. A large variety of computed descriptors in combination with numerous powerful statistical and machine learning techniques allow creating predictive and robust models with superior performance [8,9]. Nevertheless, many of these variable selection methods are linear (eliminating one descriptor at a time) and work very slowly with large datasets and a large number of descriptors [6,7]. Thus, development of fast and accurate methods for selection of a set of the most important descriptors remains a challenging problem.

Here, we introduce a fast and efficient algorithm for descriptor selection for artificial neural networks. Previously, we showed that the descriptor pruning algorithms are a very efficient way for reducing the number of input parameters and selecting a set of the most predictive ones for a variety of activities [10,11]. These algorithms operate in a manner similar to step-wise multiple linear regression analysis and on

* Corresponding author. Tel.: +380 44 573 2595; fax: +380 44 573 2552.
** Correspondence to: G. Poda, Drug Discovery Program, Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, ON M5G 0A3, Canada. Tel.: +1 647 260 7947.
E-mail addresses: vkovalishyn@yahoo.com (V. Kovalishyn), gennady.poda@oicr.on.ca (G. Poda).

http://dx.doi.org/10.1016/j.chemolab.2015.10.005
0169-7439/

each step, exclude one input parameter that was estimated to be non-significant [6]. Here we apply our previously developed methodology [12,13] for selection of input variable by ANN in combination with descriptors clustering with Kohonen's self-organizing map (SOM). We combined together our set of algorithms, which consists of the feed-forward neural networks trained with a back propagation learning algorithm (BPNN) [14], self-organizing map of Kohonen [15], *k*-nearest neighbors (KNN), and descriptor pruning methods applied earlier [6,7]. First, we used the Kohonen's SOM to perform descriptors clustering, and then we applied our pruning methods to select a subset of descriptors from each cluster for BPNN training. The combination of these two algorithms results in identification of a much smaller set of predictive descriptors in a fraction of the time when compared to the known linear selection methods. The *k*-nearest neighbors approach was used for correction of predicted values averaged over an ensemble of neural networks based on errors in prediction of *k*-nearest neighbors in chemical space or in space of an ensemble of BPNN models [16]. This procedure of correcting predicted values based on a set of "neighboring" molecules (in chemical, descriptor or model space) is targeted to diminish the systematic error for a subset of chemical space and known as local correction (LC) or associative memory approach [17,18]. The combination of BPNN and *k*-nearest neighbors for local correction is known as associative neural network (ASNN) [16]. ASNN uses correlation between ensemble responses as a measure of distance *amid* the analyzed cases for the nearest neighbor technique. The associative memory significantly improves predictive accuracy of models without a need to retrain the neural network ensemble of models.

2. Materials and methods

The study was divided into four parts: (1) creation of the training and test sets, building molecules, and optimizing their 3D geometries; (2) calculation of molecular descriptors; (3) derivation of predictive models using ANN with linear and BPA descriptor selection; and (4) validation of resulting models. We used ChemAxon software to optimize the 3D geometries of the molecules and Dragon 5.5 for descriptor calculation. Analysis of the data, descriptor selection, and model building was performed with the ASNN approach using our BPA routine written in C++.

2.1. Composition of training and test sets and calculation of molecular descriptors

We validated our new descriptor selection algorithm on two datasets from our previously published work [19,20]. The first dataset consisted of 584 compounds capable of inhibiting *M. Tuberculosis* (MTB) with activities expressed as minimal inhibitory concentrations (MIC) ranging from 0.00328 to 2800.34 μM. As the second dataset we chose a diverse set of 1015 inhibitors of phosphodiesterase type 4 (PDE4) with IC_{50} values ranging from 0.05 nM to 660 μM. The SMILES strings representing the molecular structures, their corresponding activities, and a full list of publications are available as an Excel file in the *Supplementary Materials*.

The 3D geometries of molecules were generated by ChemAxon Standardizer [21] from their SMILES notations and stored in SDF format. A set of 3224 molecular descriptors was then computed by Dragon 5.5 software [22]. Dragon calculates a large set of molecular descriptors such as geometrical, constitutional and topological descriptors, connectivity and information indices, topological charge indices, atom-centered fragments, molecular properties, and many others [1,22]. Constants and descriptors with coefficient with variance less than 5% were deleted. In addition, if any of two descriptors were at least 99% correlated, one of them was deleted. Detailed description of the Dragon descriptors can be found here: http://www.talete.mi.it/products/dragon_description.htm. As result, the 1192 descriptors were kept for the MTB dataset and 1226 for the PDE4 dataset.

2.2. Batch pruning algorithm

Our new developed algorithm consists of two main recursive parts.

Unsupervised: the self-organized map of Kohonen was used for descriptor clustering and selection of a smaller descriptor subset from each cluster.

Supervised: an ensemble of ASNN was used for building predictive models with high generalization capability.

These steps were repeated until ASNN could not be trained further in terms of decreasing the root mean square error (RMSE) of predictive values, and improving cross-validated squared correlation coefficient, Q^2. The flow chart diagram describing the developed algorithm is given in Fig. 1. Here is a more detailed description of the proposed algorithm.

During the first step, the training set was prepared, molecular descriptors were calculated, and supervised learning was performed using a variant of BPNN known as the Associative Neural Network with one hidden layer [16]. The SuperSAB algorithm was used in BPNN training [23]. The neural networks had a number of input neurons equal to the number of descriptors. One hidden layer with 5 neurons was used in the calculations (please refer to Results and discussion how we arrived at this number). Weights were initialized with different random numbers within [−0.5; +0.5] for each of 100 networks in the

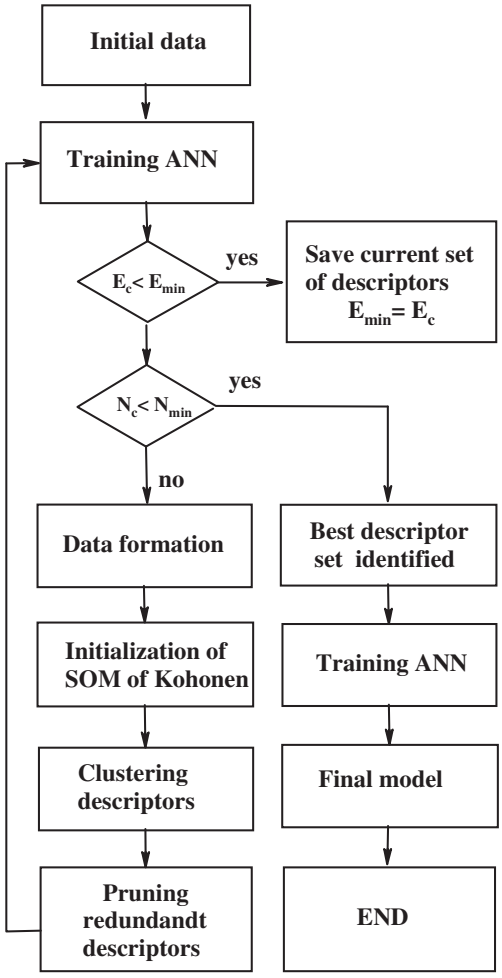

Fig. 1. Flowchart diagram for the batch pruning algorithm (BPA) variable selection procedure for artificial neural networks. N_c is the current number of descriptors, N_{min} is the minimal number of descriptors (usually, $N_{min} = 2$). E_c is the current RMSE value; E_{min} is the minimal RMSE (in the first step, $E_{min} = 10^6$).