Engineering Applications of Artificial Intelligence 71 (2018) 251–258

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Prediction of the transition temperature of bent-core liquid crystals using fuzzy "digital thermometer" model based on artificial neural networks

Davor Antanasijević [a,*], Jelena Antanasijević [b], Viktor Pocajt [b]

[a] University of Belgrade, Innovation Center of the Faculty of Technology and Metallurgy, Karnegijeva 4, 11120, Belgrade, Serbia
[b] University of Belgrade, Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia

ARTICLE INFO

Keywords:
Digital GRNN model
Prediction error estimation
SOM feature selection
QSPR

ABSTRACT

A dataset containing transition temperature values for 243 bent-core liquid crystal (LC) compounds was used to develop quantitative structure–property relationship (QSPR) models using only 2D molecular descriptors and general regression neural network (GRNN). Beside a standard analogue GRNN model, another GRNN model with fuzzy digital response was created with the aim to estimate the prediction error for each compound. Two approaches for the selection of most relevant subset of descriptors, namely the partial mutual information (PMI) and self-organizing maps combined with chi square ranking, were also compared. The best results were obtained using analogue GRNN model based on PMI selected subset ($R^2 = 0.91$), with the mean absolute error (MAE) lower in comparison with previously published corresponding QSPR models. The digital PMI-GRNN model enabled distinction between high and low accurate predictions, i.e. ones with absolute error higher than mean absolute error (MAE) and others with absolute error ≤ MAE, with the accuracy of 81%.

1. Introduction

Quantitative structure–property/activity relationship (QSPR/QSAR) models for the prediction of physical properties and biological activities of organic compounds from their molecular structures have been a focus of great attention for a long time (Mu et al., 2007). In recent years, artificial neural networks (ANNs) have opened a new avenue for developing empirical models useful for predicting the performance of the process outside the experimental domain (Olea, 2007). A literature review indicate that QSPR modeling based on ANNs has grown dramatically, suggesting the importance of applications of ANN in molecular modeling (Katritzky et al., 2010). ANNs are useful tools in QSPR studies because a given structure–property relationship is often nonlinear (Yao et al., 2004).

The QSPR models are very important in the field of liquid crystals, since small structural modifications of liquid crystal (LC) molecules can drastically influence their transition temperature. Therefore, the design of molecular structure with desired LC phase temperature, based only on empirical rules, is a very complex task. The QSPR methodology has been successfully applied to predict various physical and chemical properties of LCs (Al-Fahemi, 2014; Antanasijević et al., 2016a; Gong et al., 2008; Johnson and Jurs, 1999), which possess unique physicochemical properties and have wide application in a variety of fields (Bahadur, 1994; Vicari, 2016). Nowadays, bent-core LCs are the most attractive type of these materials due to their promising applications (Eremin and Jákli, 2013; Takezoe and Takanishi, 2006), and their transition temperatures have been the subject of recent QSPR studies (Antanasijević et al., 2016b, c). In those studies, the QSPR models with good predictive ability ($R^2 \geq 0.90$) based on decision trees, multivariate adaptive regression splines and group method of data handling (GMDH-type) neural network have been proposed for the prediction of five-ring bent-core LCs transition temperatures.

In this study, the transition temperature of bent-core LCs was predicted using general regression neural network (GRNN) with the aim to develop a more accurate model that will allow the estimation of prediction error for each compound. For this purpose, the ANN model with digital output neurons, previously proposed for the transition temperatures of smectic LC compounds by Schroder et al. (1996), has been adapted using fuzzy logic. Since a large number of molecular descriptors can be calculated for each compound, two model-free methods that can take into account both linear and nonlinear relationships, namely the partial mutual information (PMI) and self-organizing maps (SOM) combined with chi square ranking, were applied prior to the model development for the selection of the most relevant subset.

* Corresponding author.
E-mail address: dantanasijević@tmf.bg.ac.rs (D. Antanasijević).

https://doi.org/10.1016/j.engappai.2018.03.009
Received 25 June 2017; Received in revised form 5 December 2017; Accepted 12 March 2018

2. Material and methods

2.1. LC dataset

In this study, a dataset (see Table S1 in the supplemental material of the paper (Antanasijević et al., 2016c) that contains the transition temperature values for 243 bent-core LC compounds was utilized for the development and testing of models. This dataset consisted of structurally diverse five-ring aromatic compounds with the transition temperature values in the range from 352.15 to 458.15 K. The same subset of 36 compounds was used for model testing, in order to allow direct comparison with the models created in the our previous studies (Antanasijević et al., 2016b, c).

2.2. Descriptors and their selection

After the molecular structures were sketched in ChemDraw, the calculation of 2D molecular descriptors was performed using PaDEL-Descriptor software (Yap, 2011). Although PaDEL calculates more than 600 2D descriptors, the subsequent elimination of constant and near constant descriptors has yield 360 descriptors that were further used. Concerning the importance of dimension reduction (Ma and Zhu, 2013) in QSPR modeling, two approaches were compared:

(a) PMI-based (partial mutual information), which is a non-linear input selection technique that is considered as highly suitable for development of ANN models (May et al., 2008), and
(b) feature selection – self organizing map (FSL-SOM) that is proposed in this study.

PMI is used because of its abilities to take into account linear and non-linear input–output relationships and to determine the significance of selected inputs (Li et al., 2015). PMI uses a partial measure of the mutual information criterion (MIC) in order to determine the inputs that have highly significant relationship with the output variable of the system being modeled (Kalteh et al., 2008). The MIC is a measure of dependence between any two variables (e.g. X and Y), and it is defined as (Sharma, 2000):

$$MIC = \iint f_{X,Y}(x,y)\, log_e \left[\frac{f_{X,Y}(x,y)}{f_X(x)\, f_Y(y)}\right] dx dy \quad (1)$$

where $f_X(x)$ and $f_Y(y)$ are the marginal probability density function's (PDF's) of X and Y, respectively, and $f_{XY}(x, y)$ is the joint (bivariate) PDF of X and Y. From MIC the PMI between X as independent and Y as dependent variable for a set of pre-existing descriptors Z, can be calculated as numerical estimation:

$$PMI = \frac{1}{n}\sum_{i=1}^{n} log_e \left[\frac{f_{X',Y'}(x'_i, y'_i)}{f_{X'}(x'_i)\, f_{Y'}(y'_i)}\right] \quad (2)$$

where x'_i and y'_i are the ith residuals in the sample data set of size n, while corresponding marginal and joint probability densities are determined for $X' = X - E[X|Z]$ and $Y' = Y - E[Y|Z]$, where E denotes the expectation operation.

Among several PMI termination criterions (May et al., 2008), the Hampel score (Z), which is an outlier test applied to the descriptor with the highest PMI value, relative to the distribution of PMI values for all descriptors, was applied in this study. The Hampel distance test (Z_j), i.e. modified Z-score, begins by calculating the absolute deviation (d_j) from the median PMI for all descriptors, after Hampel distance is calculated using the following formula (May et al., 2008):

$$Z_j = \frac{d_j}{1.4826 MAD} \quad (3)$$

where MAD denotes the median absolute deviation. The factor of 1.4826 scales the distance, so the rule $Z > 3$ can be applied for the selection of most significant set of descriptors.

FLS-SOM can be regarded as an adjustment of previously proposed SOM-GAGRNN (genetic algorithm general regression neural network) by Bowden et al. (2005) in order to obtain a model-free approach. As in case of Bowden et al. (2005) approach, the SOM is firstly applied to cluster the descriptors into groups of similar inputs, but instead of using the GRNN model to select the most significant inputs, we have applied the feature selection routine from Statistica (StatSoft. Inc., 2010) that ranks the molecular descriptors according to their statistical association with the modeled output using a chi square based ranking (Newby et al., 2013), in order to determine the most important descriptor from each cluster (Fig. 1). The advantage of the SOM is that it is able to take into account nonlinear cross-dependence between inputs (Bowden et al., 2005).

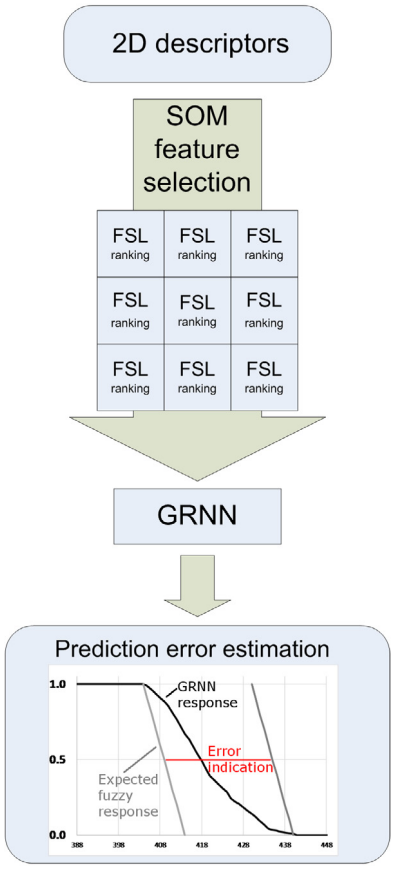

Fig. 1. Schematic representation of FSL-SOM-GRNN modeling, with the estimation of prediction error based on fuzzy digital response.

SOM is a two layer unsupervised classifier, where the size of the first (input) layer is determined by the number of inputs while the size of the second (output) layer must be determined empirically. Therefore, in order to determine the optimal number of output map units, i.e. the number of descriptors further used in modeling, we have started from a low size map (3 × 3) and gradually increased their size (4 × 4, 5 × 5 etc.) until at least one map unit have remained "empty" after SOM converged. FSL is then applied on each map unit, and the set of descriptors with limited cross-dependence is determined.

2.3. GRNN architecture and training

General regression neural network (GRNN) was invented by Specht (Specht, 1991) as a modification of probabilistic neural network for the prediction of continuous outputs. In numerous studies, it was found that GRNN responds much better than ANNs trained with standard back-propagation algorithm to many types of problems, although this is not a rule (Kalogirou, 2003). GRNN is a one-pass four-layer feed-forward ANN that consists of the input layer (I neurons in Fig. 2), the pattern layer