IAU-ARAK

J. Iran. Chem. Res. 4 (2011) 97-103

Journal of the Iranian Chemical Research

www.iau-jicr.com

Prediction of melting points of a diverse chemical set using fuzzy regression tree

Vali Zare-Shahabadi*

Department of Chemistry, Mahshahr Branch, Islamic Azad University, Mahshahr, Iran

Received 8 February 2011; received in revised form 15 March 2011; accepted 16 March 2011

Abstract

The classification and regression trees (CART) possess the advantage of being able to handle large data sets and yield readily interpretable models. In spite to these advantages, they are also recognized as highly unstable classifiers with respect to minor perturbations in the training data. In the other words methods present high variance. Fuzzy logic brings in an improvement in these aspects due to the elasticity of fuzzy sets formalism. ACS, which is a meta-heuristic algorithm and derived from the observation of real ants, was used to optimize fuzzy parameters. The purpose of this study was to explore the use of fuzzy regression tree (RT) for modeling of melting points of a large variety of chemical compounds. To test the ability of the resulted tree, a set of approximately 4173 structures and their melting points were used (3000 compounds as training set and 1173 as validation set). Further, an external test set contains of 277 drugs were used to validate the prediction ability of the tree. Comparison the results obtained from both trees showed that the fuzzy RT performs better than that produced by recursive partitioning procedure.

Keywords: Ant colony system; Classification; Regression tree; Melting points.

1. Introduction

Decision tree (DT) techniques belong to computational intelligent methods and became highly popular in the age of modern computers. They are based on a sequence of questions that can be answered by either yes or no. Each question queries whether a predictor satisfies a given condition, whereby the condition can be both continuous and discrete. Depending on the answer to each question, one can either proceed to another question or arrive at a response value. DTs can be used for non-linear regression (Regression Tree) when using continuous variables or they can be used for classification (Classification Tree) when using discrete classes. When used for feature selection, the DT is grown as a regression tree [1].

Without prior knowledge of the nonlinearity, the regression tree is capable of approximating any non-linear relationship using a set of linear models. Although regression trees are interpretable representations of a non-linear input-output relationship, the discontinuity at the decision boundaries is unnatural and brings undesired effects to the overall regression and generalization of the problem [2]. CART offers several advantages over alternative quantitative structure activity/property relationship (QSAR/QSPR) methodologies, including simplicity,

* Corresponding author. Tel. & fax: +98 7117402815.
E-mail address: valizare@gmail.com (V. Zare-Shahabadi)

interpretability, and the ability to handle large data sets [3]. In addition, another major advantage is that it makes no underlying assumptions regarding the distribution of the values of the predictors. It means that the tree is constructed in a recursive binary way, resulting in nodes connected by branches.

Although decision tree techniques have already been shown to be interpretable, efficient, problem independent and able to treat large scale applications, but they are also recognized as highly unstable classifiers with respect to minor perturbations in the training data. In the other words methods present high variance. Fuzzy logic brings in an improvement in these aspects due to the elasticity of fuzzy sets formalism [2].

In this work, ant colony optimization (ACO) [4, 5] was used to build a fuzzy regression tree (RT) and the resulted tree was used for quantitative structure-property relationship study of the melting points of 4173 structurally diverse chemicals.

The melting point is not only important in the screening for solid-state characteristics, but also is used as a descriptor for predicting other properties, such as solubility [6-8]. However, the experimental estimation of melting points is believed to be a difficult task [9]; many compounds, including some environmentally and pharmaceutically important ones, decompose prior to or during melting. In addition, the estimation of melting points from structural properties (QSPR) is generally more difficult than the prediction of some other properties such as boiling points [10].

The influence of different molecular parameters on melting points has already been the subject of intense research [10-17]. A number of these studies focused on very restricted classes of compounds [13], or developed a non-linear model to predict melting points [10, 17]. Karthikeyan et al. [10] proposed a nonlinear artificial neural network (ANN) model for predicting the melting point of a diverse data set of 4173 compounds and. This model achieved a maximum correlation coefficient of $R^2 = 0.662$. The goal of this work was development of a predictive model based on fuzzy RT for the melting point of the data set collected by Karthikeyan et al. [10].

2. Data set

The data set consisted of melting points of 4173 chemicals, used by Karthikeyan et al. [10]. The molecules were partitioned into training and validation sets, with respective sizes of 3000 and 1173, by sample set partitioning based on joint x-y distances (SPXY) method [18].

A large number of 2D and 3D descriptors to capture molecular, physicochemical and other graph-based properties were employed, calculated by Karthikeyan et al. [10]. The 2D descriptors include physical properties (such as charge, van der Waals volume, and molecular refractivity), subdivided surface areas (atomic contributions to logP and molecular refractivity), counts of elemental atom types and of bond types, Kier/ Hall connectivity and kappa shape indices, topological indices (Wiener index and Balaban index), pharmacophore feature counts (number of acidic and basic groups and hydrogen bond donors and acceptors), and partial charge descriptors.

The conformation-independent 3D descriptors include potential energy terms (such as total potential energy and contributions of angle bend, electrostatic, out-of-plane, solvation, etc. terms) and surface area, volume, and shape descriptors (among them water accessible surface area, mass density, and principal moments of inertia). A total number of 202 descriptors were finally used in this study.

3. Theory of ant colony system

Dorigo and his colleagues were the first to apply ant colony, referred as ant system (AS), to the traveling salesman problem [19]. Later on, a more promising method, named as ant colony system (ACS) was developed [20]. Previously, we employed both algorithms to solve variable selection problem in multivariate calibration and QSPR studies [21].

In this study, ACS algorithm was used to build a fuzzy regression tree (RT). First of all, a RT based on the algorithm proposed by Breiman was built [1]. To obtain a fuzzy RT, a fuzzy membership function with proper parameters should be applied at each node. ACS was used to optimize the fuzzy parameters of the RT. It is worthy to note that variables used to split molecules at each node were those selected previously by Breiman's algorithm and the ACS algorithm was used to optimize the value of cut point at each node and the upper and lower thresholds.

The ACS-CART algorithm [22] is briefly described in the following: supposed the optimum RT obtained by Breiman's algorithm has *m* nodes. At each node, an appropriate cut point must be selected for the chosen variable. Selection is done based on cut-point pheromone matrix. For each variable, ten cut points between minimum and maximum quantity of each variable in data matrix are chosen. It is suggested that a variable at each node has different optimum cut point and because that there are *m* nodes, cut points pheromone matrix must be a matrix of size $m\times 10$. At the beginning, an equal amount of pheromone values are assigned to each cut points ($\tau_{i,j}$). Based on this pheromone matrix a probability vector is calculated for the ith selected variable.

To divide members of a parent node into children nodes a membership function is required. Preliminary studies showed that from four types of membership functions (MFs), Triangular-shaped MF (TMF) was the best. Other studied MFs were: Gaussian MF, Generalized bell-shaped MF, and Sigmoidally shaped MF. TMF has three parameters: *a*,*b*, and *c*, where *a* and *c* locate the "feet" of the triangle and the parameter *b* locates the peak. The value of the cut-point is the *b* value of the TMF. The remaining two parameters (*a* and *c*) are selected such that a symmetric Triangular shape is produced.

As mentioned above, there is a cut-points pheromone matrixes based on which an ant selects a cut-point value at each node. Details presented in our previous paper [5].When a colony of ants is created, all ants are evaluated as followed: for an ant, in each terminal node a weighted MLR model with selected descriptors by the stepwise method is developed and calculated coefficients are stored (recall that each ant is a fuzzy regression tree). Weights in the weighted MLR are equal to the membership value of each molecule in the terminal node. Note that membership function at each node is a curve that defines how each molecule in the parent node is mapped to a membership value (or degree of membership) between 0 and 1. So, summation of the membership values of a molecule in all terminal nodes equal to unity. After that, the best ant, ant with highest correlation coefficient in the regression model, updating both pheromone matrixes, *i.e.,* the pheromone values of selected cut-point by the best ant is increased by a constant factor υ (setting to 0.4 by trial and error):

$$\tau_{i,k}(t+1)=\tau_{i,k}(t)+\lfloor\tau_{i,k}(t)\times\upsilon\rfloor \qquad k\in\beta$$

where β is set of selected variable by the best ant, *k* is each of selected variable, *t* is number of iteration and *i* indicates number of node.

At the end of each of iteration, pheromone evaporation is done as followed:

$$\tau_{i,k}(t+1)=\tau_{i,k}(t)-\lfloor\tau_{i,k}(t)\times\rho\rfloor$$

Therefore, after the first iteration, the best cut-point values that used to split compounds at each node by the best ant give higher pheromone values than the others.

4. Results and discussion

The goal of this work is development of a predictive model to estimate melting point of various chemicals. To obtain such a model, a diverse set of chemicals containing 4173 molecules belonging to a wide variety of organic compounds was used. Fig 1 represents the histogram of the melting point data, indicating that the data have a nearly normal distribution. A large standard deviation in the melting point data (i.e., 65 °C) indicates diversity of the used data points.

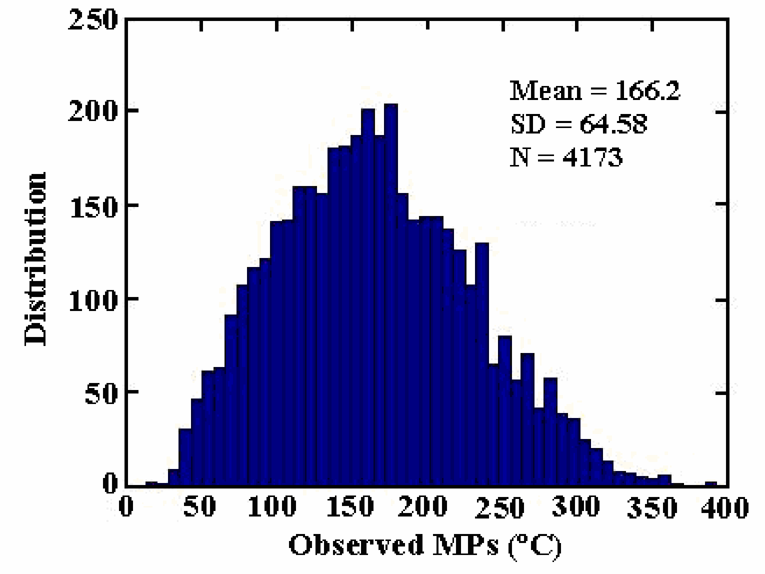

Fig. 1. Histogram of the distribution of the experimental melting points for the total data set of 4173 chemicals.

In attempting to develop a structure-melting point relationship with sufficient predictive ability, the RT analysis that represented good predictive ability in QSAR/QSPR studies [3] was used. First of all, an ordinary regression tree, obtained by the algorithm proposed by Breiman [1], was constructed. This tree, after pruning stage, classified 4173 compounds of training and validation sets to 41 groups (shown in Fig. 2). It represented poor prediction ability with correlation coefficients of 0.41 and 0.11 for the training and validation sets, respectively. Descriptors used to split root node up to node No. 7 are TPSA, balabanJ, a_nN, PM3_LUMO, MNDO_LUMO, b_1rotR, and vsa_acc. These descriptors are explained in Table 1.

Table 1
Descriptors used in constructing ordinary CART as split criteria.

Descriptor	Definition
TPSA	Polar surface area calculated using group contributions to approximate the polar surface area from connection table information only.
balabanJ	Balaban's connectivity topological index
a_nN	Number of nitrogen atoms
PM3-LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian [MOPAC].
MNDO-LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the MNDO Hamiltonian [MOPAC].
b-1rotR	Fraction of rotatable single bonds
vsa-acc	Approximation to the sum of VDW surface areas of pure hydrogen bond acceptors

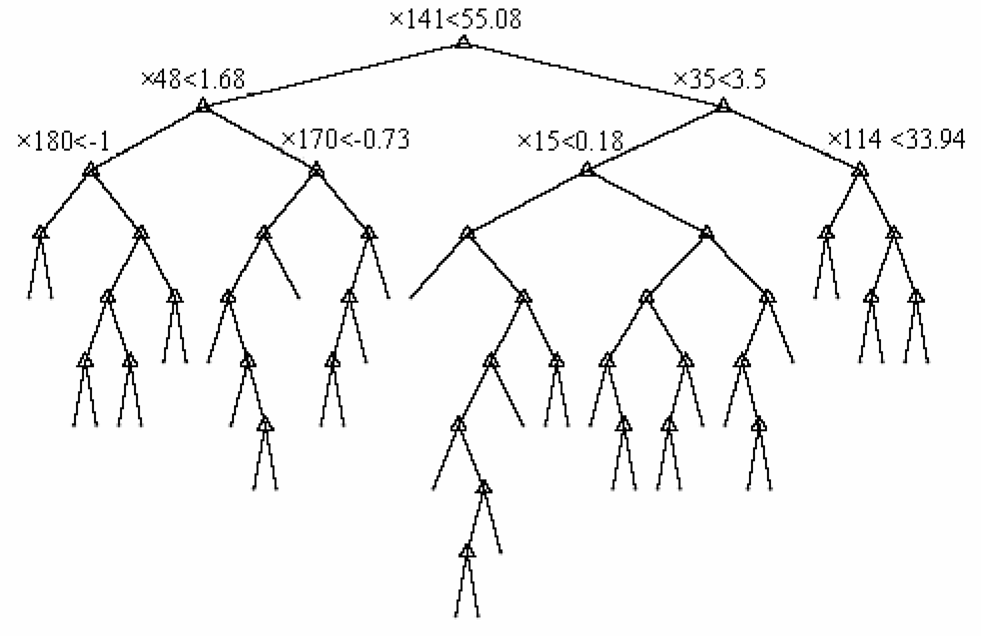

Fig. 2. Tree obtained by ordinary CART algorithm.

One of the reasons for poor prediction ability of the tree is that it uses mean of each terminal nodes as a prediction value for all of its members. Therefore, in the first step, to improve the results of the tree, a multivariate linear regression model (MLR) is built in each terminal node and the calculated coefficients were used to predict the melting points of validation molecules. In each terminal node, the proper descriptors were selected by stepwise variable selection method [22]. The root mean square errors (RMSE) values for the training and validation sets are 36.93 and 109.81 °C, respectively. The respective R^2 values for the training and prediction sets are 0.56 and 0.12. High value of R^2 for training set together with low value for prediction set indicates the presence of over-fitting in the resulted model.

In the consequent step, we decided to combine the ACS algorithm with CART algorithm to overcome the disadvantage of the CART by applying the fuzzy role at each node of the tree.

As noted previously, ACS algorithm has some parameters which should be optimized to obtain better results. These parameters include ϕ, υ, ρ, and number of iterations. All of these parameters, except ϕ and number of iterations, were set to optimum value by trial and error. The optimum values were 0.4 and 0.4, for υ and ρ, respectively. The proper values for ϕ and number of iterations were found to be 30 for both parameters. It is worth mentioning that larger values for these parameters, higher the chance of finding the global optima, together with an expected increase in computation time.

The obtained fuzzy tree revealed better prediction ability over the conventional CART model (i.e., R^2 of training set of 0.72 and of prediction set of 0.79). Note that at each terminal node, a MLR model with those variables selected by stepwise method was developed. A plot of predicted versus observed melting points exhibits good correlation for the majority of compounds. The observed improvement is obviously the results of applying the fuzzy role on the CART.

To further investigate the predictive ability of the model, it was used to predict the melting points of an external validation set of 277 drugs. The results are also including in Fig. 3. The high prediction ability is observed from this figure (R^2 = 0.78),

As it is seen from Fig. 3, some compounds exhibit significant deviation from the regression line. These compounds belong to two groups: small molecules with self-organizing ability in the solid state that results in stronger interactions than ordinary molecules of the same size and larger molecules whose physicochemically similar neighbors are just in an area of chemical space that does not allow for appropriate interactions result in underestimated melting points [10].

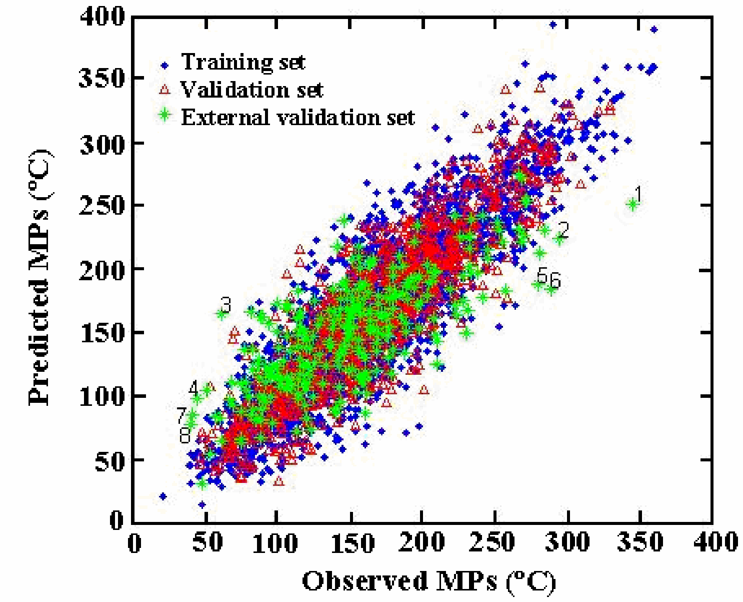

Fig. 3. Plot obtained for a training set (filled blue diamonds), a validation set (unfilled triangles), and an external validation set (stars) between experimental and predicted melting points.

5. Conclusion

Regression tree is a powerful method in data mining filed. It can handle large data set and develop linear and/or non-linear relationship between independent variables and dependent variable. However, RT has some disadvantages. For example, it has high variance. In this work, the ACS algorithm was used to build a fuzzy regression tree which has higher stability compared to the ordinary RT. The resulted fuzzy tree was successfully employed for prediction of melting points of a large set of chemicals.

Acknowledgment

This paper was extracted from a research project entitled "Applying ACO algorithm to fuzzyfication of regression tree and its application to construction a QSPR model for predicting melting point. The author would like to thank Dr. Fatemeh Abbasitabar for helpful discussions.

References

[1] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, 1984.
[2] R. Jang, Neuro-Fuzzy and Soft Computing, Prentice Hall, NJ, 1997.
[3] S. Izrailev, D. Agrafiotis, J. Chem. Inf. Comput. Sci. 41 (2001) 176-180.
[4] V. Zare-Shahabadi, F. Abbasitabar, J. Compt. Chem. 31 (2010) 2354-2362.
[5] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemometrics 20 (2006) 146-157.
[6] S.H. Yalkowsky, S.C. Valvani, J. Pharm. Sci. 69 (1980) 912-922.
[7] S.H. Yalkowsky, J. Pharm. Sci. 70 (1981) 971-973.
[8] Y. Ran, S.H. Yalkowsky, J. Chem. Inf. Comput. Sci. 41 (2001) 354-357.
[9] A. Gavezzotti, J.Chem. Soc., Perkin Trans. 2 (1995) 1399-1404.
[10] M. Karthikeyan, R.C. Glen, A. Bender, J. Chem. Inf. Model. 45 (2005) 581-590.
[11] J.C. Dearden, Sci. Total Environ. 109/110 (1991) 59-68.
[12] M. Charton, B. Charton, J. Phys. Org. Chem. 7 (1994) 196-206.
[13] A.R. Katritzky, U. Maran, M. Karelson, V.S. Lobanov, J. Chem. Inf. Comput. Sci. 37 (1997) 913-919.
[14] M. Charton, J. Comput.-Aided Mol. Des. 17 (2003) 197-209.

[15] C. A. Bergstrom, U. Norinder, K. Luthman, P. Artursson, J. Chem. Inf. Comput. Sci. 43 (2003) 1177-1185.

[16] L. Ma, C. Cheng, J. Chemom. 16 (2002) 75-80.

[17] K. J. Burch, E. G. Whitehead, J. Chem. Eng. Data 49 (2004) 858-863.

[18] R.K.H. Galvão, M.C.U. Araujo, G. E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Talanta 67 (2005) 736-740.

[19] M. Dorigo, Optimization, Learning and Natural Algorithms, Ph.D. Thesis, Politecnico di Milano, Italy, 1992.

[20] M. Dorigo, T. Stutzle, Ant Colony Optimization, MIT Press, New York, 2004.

[21] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, QSAR & Comb. Sci. 28 (2009) 1263-1275.

[22] V. Zare-Shahabadi, Ant colony optimization and its applications in analytical chemistry, Ph.D. Thesis, Shiraz University, Shiraz, Iran, 2008.