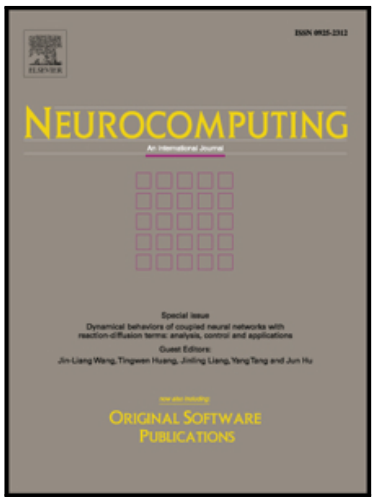

Optimization Enhanced Genetic Algorithm-Support Vector Regression for the Prediction of Compound Retention Indices in Gas Chromatography

Jun Zhang , Chun-Hou Zheng , Yi Xia , Bing Wang , Peng Chen

PII: S0925-2312(17)30312-0
DOI: 10.1016/j.neucom.2016.11.070
Reference: NEUCOM 18108

To appear in: *Neurocomputing*

Received date: 28 December 2015
Revised date: 15 October 2016
Accepted date: 22 November 2016

Please cite this article as: Jun Zhang , Chun-Hou Zheng , Yi Xia , Bing Wang , Peng Chen , Optimization Enhanced Genetic Algorithm-Support Vector Regression for the Prediction of Compound Retention Indices in Gas Chromatography, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.11.070

Optimization Enhanced Genetic Algorithm-Support Vector Regression for the Prediction of Compound Retention Indices in Gas Chromatography

Jun Zhang[a], Chun-Hou Zheng[a], Yi Xia[a], Bing Wang[b], Peng Chen[c*]

[a]School of Electronic Engineering & Automation, Anhui University, Hefei, Anhui,230601, China
[b]School of Electrical and Information Engineering, Anhui University of Technology,Ma Anshan, 243032, China
[c]Institute of Health Sciences, Anhui University, Hefei, Anhui, 230601,China

Abstract

A new method using genetic algorithm and support vector regression with parameter optimization (GA-SVR-PO) was developed for the prediction of compound retention indices (RI) in gas chromatography. The dataset used in this work consists of 252 compounds extracted from the Molecular Operating Environment (MOE) boiling point database. Molecular descriptors were calculated by descriptor tools of the MOE software package. After removing redundant descriptors, 151 descriptors were obtained for each compound. A genetic algorithm (GA) was used to select the best subset of molecular descriptors and the best parameters of SVR to optimize the prediction performance of compound retention indices. A 10-fold cross-validation method was used to evaluate the prediction performance. We compared the performance of our proposed model with three existing methods: GA coupled with multiple linear regression (GA-MLR), the subset selected by GA-MLR used to train SVR (GA-MLR-SVR), and GA on SVR (GA-SVR). The experimental results demonstrate that our proposed GA-SVR-PO model has better predictive performance than other existing models with $R^2 > 0.967$ and RMSE =49.94. The prediction accuracy of GA-SVR-PO model is 96% at 10% of prediction variation.

Keywords: support vector regression, quantitative structure-retention relationship, retention indices prediction, gas chromatography

*Corresponding author Tel:+86-13855180174
Email address:wwwzhangjun@163.com