Chemometric versus Random Forest Predictors of Ionic Liquid Toxicity

Ž. Kurtanjek
University of Zagreb, Faculty of Food Technology and Biotechnology,
Pierottijeva 6, 10000 Zagreb, Croatia

doi: 10.15255/CABEQ.2014.19399

Original scientific paper
Received: July 1, 2014
Accepted: November 28, 2014

This manuscript is dedicated to Prof. Egon Bauman (90)

The objective of this work was a comparative analysis of the standard chemometric and decision tree(s) models for prediction of biological impact of ionic liquids (ILs) for various combinations of cations and anions. The models are based on molecular descriptors for combinations of the following cations: imidazole, pyridinium, quinolinium, ammonium, phosphonium; and anions: BF_4, Cl, PF_6, Br, CFNOS, NCN_2, $C_6F_{18}PBF_4$, $C_6F_{18}P$. The derived data matrix is decomposed by singular value decomposition of the cation and anion matrices into corresponding first ten components, each accounting for 99.5 % of the corresponding total variances. Biological impact data, i.e. molecular level toxicity, are based on acetylcholinestarase inhibition experimental data provided in MERCK Ionic Liquids Biological Effects Database. Applied were the following models: Principal component regression (PCR), partial least squares (PLS), and decision tree(s) model. The model performances were compared by ten-fold validation. Obtained were the following Pearson regression coefficients R^2: PCR 0.62, PLS 0.64, and for decision tree forest RFDT 0.992. The decision tree(s) models significantly outperformed chemometric models for numerical predictions of EC_{50} concentrations and the classification of ILs into four levels of toxicities.

Key words:
ionic liquids, toxicity, chemometrics, decision tree

Introduction

Interest in ionic liquids ILs stems from their unique solvent properties and potential process "self-containment". Their application in chemical processes and biotransformations provides the possibility for clean manufacturing ("green technology"). Besides their solvent and extraction functions, ILs also exhibit synergy effects with catalysts (enzymes) yielding higher production productivity. Theoretically, there is a limitless number of possible ILs with a very broad range of physical and chemical properties. Research on ILs has become one of the most interesting application research areas in novel catalytic synthesis, biofuel production from agricultural wastes, integration of chemical and enzyme reactors with separation processes, polymerization, nanotechnology, enzyme-catalysis, composite preparation and renewable resource utilization[1–3]. Especially interesting is the use of microreactors for ionic liquid synthesis and possibly as production systems for integrated biotransformations and product separation[4]. However, the recent questions of ILs' eco-toxicity and their degradability have also been raised.

Analysis of their versatile structure is formally viewed as a combinatorial problem which can be effectively accounted by computers. The object of this work is to apply computer modeling by chemometric methodology and decision tree algorithm for predicting continuous variables, such as toxicity level concentration EC_{50} and level classification, based on the choice of cation and anion structure and their chemical compositions. Predictions of ILs physical properties are based on literature published data and internet available NIST and MERCK databases of physical properties and cytotoxicity[5–7].

The main objective of this work is in inferring the rules and patterns implicitly contained in a set of chemical structures and molecular descriptors. Applied is a supervised learning algorithm with target sets for continuous and classification properties revealing relationships between molecular descriptors.

Experimental

The chemical formula of each ion is recorded in SMILES and MOL format and evaluated for corresponding molecular descriptors[8]. Hence, each combination of ions for a specific IL is represented

* Corresponding author: zkurt@pbf.hr

by 2x797 data points. Since numerical values of molecular descriptors cover a range of numerical orders of magnitude, each descriptor is autoscaled based on the sample average and the corresponding standard deviation. For the selected cations, the transformation is:

$$\mathbf{X}_C \leftarrow \frac{\mathbf{X}_C - \bar{\mathbf{X}}_C}{\sigma(\mathbf{X}_C)} \tag{1}$$

Similarly, molecular descriptors for the selected anions are transformed accordingly:

$$\leftarrow \frac{\mathbf{X}_A \quad \bar{\mathbf{X}}_A}{(\quad)} \tag{2}$$

The obtained matrices of the autoscaled descriptors are analyzed for their mutual inter-relationships. For cations and anions data matrices, the average Pearson correlation of R^2 =0.4 is obtained, which is significant considering the large number of samples (ions). Due to high co-linearity between various molecular descriptors, the data matrices are decomposed into a series of partial components by application of singular value decomposition of the corresponding anion $\mathbf{X}_A$ and cation $\mathbf{X}_C$ covariances by solving the eigenvalue problems:

$$\left(\mathbf{X}_C^T\mathbf{X}_C\right)\mathbf{v}_{C,i} = \lambda_{C,i}\,\mathbf{v}_{C,i} \quad j = 1,2\cdots M \tag{3}$$

$$\left(\mathbf{X}_A^T\mathbf{X}_A\right)\mathbf{v}_{A,i} = \lambda_{A,i}\,\mathbf{v}_{A,i} \quad j = 1,2\cdots M \tag{4}$$

Decomposed matrices, $\mathbf{P}_A$ and $\mathbf{P}_C$, are defined by the corresponding eigenvectors $\mathbf{v}_A$ and $\mathbf{v}_C$, and the contributions of individual partial decompositions are evaluated by the ratios of squares corresponding eigenvalues λ_{Ai} and λ_A to the number of descriptors *M*.

$$\mathbf{P}_C = \mathbf{X}_C \cdot \left(\mathbf{v}_{C,1}\middle|\mathbf{v}_{C,2}\middle|\cdots\mathbf{v}_{C,K}\right) \tag{5}$$

$$\mathbf{P}_A = \mathbf{X}_A \cdot \left(\mathbf{v}_{A,1}\middle|\mathbf{v}_{A,2}\middle|\cdots\mathbf{v}_{A,K}\right) \tag{6}$$

Based on the preselected level of 99.5 % of the total variance, the first ten, K=10, eigenvectors for each data set are chosen.

Results and discussion

Compared are the chemometric and decision tree models for regression and prediction of concentration E_{50} and toxicity classification for inhibition of acetylchlorinestarase inhibition experimental data provided in MERCK Ionic Liquids Biological Effects Database[7]. The model input data are the target values of molecular descriptor projections. The chemometric models are linear models, and applied here based on their expected robustness and improved prediction when compared to classical least squares multivariate models[9–12]. The first tested model is Principal Component Regression (PCR) given by Eq. (7).

$$\mathbf{Y} = \boldsymbol{\beta}_A \cdot \mathbf{P}_A + \boldsymbol{\beta}_C \cdot \mathbf{P}_C + \mathbf{E} \tag{7}$$

The statistical evaluation and analysis of the model parameters are performed by the algorithms provided by R open source software[16] and STATISTICA[17]. Applied is ten-fold cross validation within the training set of samples, as well as validation with the data set that had not been used during the modelling phase. The model "quality" for prediction of E_{50} concentration is relatively "poor" with R^2 = 0.62 presented in Fig. 1.

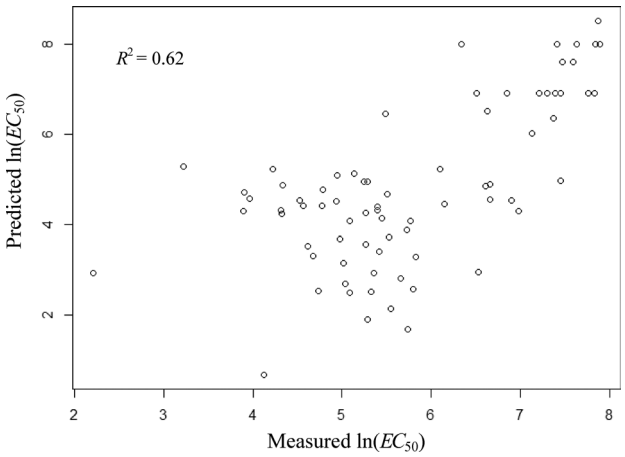

Fig. 1 – *Comparison between the test samples for measured ln(EC_{50}) concentrations and the principal component regression model predictions*

The second tested model is Partial Least Squares (PLS) which is to improve the predictions by separate decomposition of the input and output data sets (Eqs. 8-9).

$$\mathbf{X} = \mathbf{T} \cdot \left(\mathbf{P}_C\middle|\mathbf{P}_A\right)^T + \mathbf{E}_X \tag{8}$$

$$\mathbf{Y} = \mathbf{U} \cdot (\mathbf{Q})^T + \mathbf{E}_Y \tag{9}$$

The predictive model is built by regression between the inner projections **T** and **U**:

$$\mathbf{U} = \mathbf{T} \cdot \boldsymbol{\beta} + \mathbf{E} \tag{10}$$

The PLS model predictions on the test data slightly improved yielding R^2 = 0.64 as presented in Fig. 2.

The obtained relatively poor predictions of EC_{50} by the chemometric models is in contrast to good predictions for some of ILs physical proper-

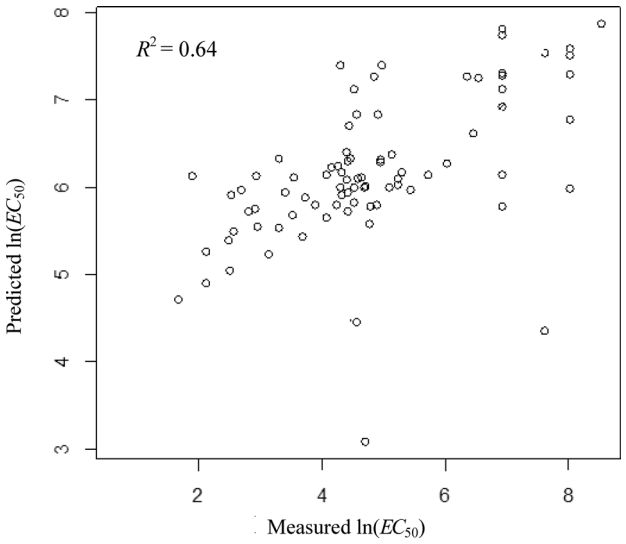

Fig. 2 – *Comparison between the test samples for measured ln(EC_{50}) concentrations and the partial least squares model predictions*

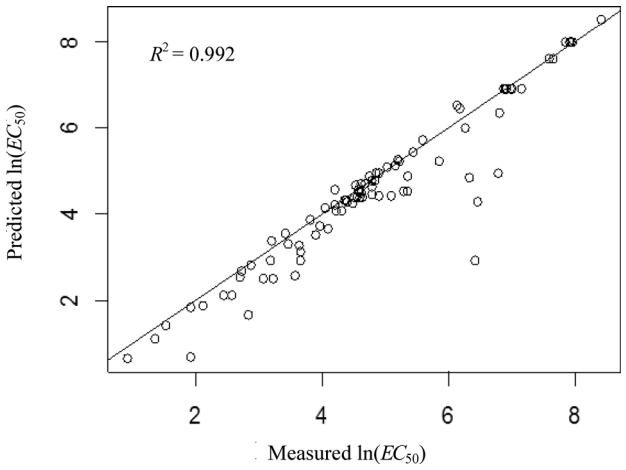

Fig. 3 – *Comparison between the test samples for measured ln(EC_{50}) concentrations and the random forest model predictions*

ties, as for example, viscosity, given in literature[12–14]. A possible explanation is due to high dispersion of the experimental data EC_{50} involved in measurement of the biological effect of ILs.

In order to elevate the modelling assumption on continuity and linearity between molecular descriptors and biological effects, applied are decision tree (DT) and random forest (RF) models[11,15,18]. These are nonparametric models and are not based on assumed functional relationships between the input and output data. The main objective of decision tree model is a supervised procedure of step-wise classification of input data by binary split into subsets for "improved" or more significant information content (information gain). It is obtained by minimisation of Gini index or pattern entropy. Produced models are not given in a closed mathematical form, but as a set of logical statements which can be easily represented in graphical form as a tree of stepwise decisions. When a DT model is used for regression, the numerical range of output data is approximated by pseudo classes for assumed precision of regression predictions. Here is applied the Breiman and Cutler[15] algorithm available in R software system and tree plotting[16–18].

$$\hat{\mathbf{Y}}_{DT}^{P} = DT\left(\mathbf{Y}, \mathbf{P}_C, \mathbf{P}_A\right) \tag{11}$$

Single decision tree prediction models tend to be biased but modelling can be improved by re-initialization of collections of trees by randomisation of the split algorithm and production of a random forest. Prediction of a random forest is obtained by aggregation of individual trees with weighted response corresponding to individual tree cross-validation.

$$\hat{\mathbf{Y}}_{RF}^{P} = RF\left(\mathbf{Y}, \mathbf{P}_C, \mathbf{P}_A\right) \tag{12}$$

Modelling results are presented in Figs. 3–4. Prediction of ln(EC_{50}) by the random forest model is greatly improved with Pearson correlation $R^2 = 0.992$. Individual decision tree for classification of ILs toxicity is depicted in Fig. 5. For acetycholinesterase the following classes were here adopted: low (L, $EC_{50} < 10$ μmol L^{-1}), medium (M), $EC_{50} \in [10 - 100$ μmol $L^{-1}]$, high (H) $EC_{50} \in [100 - 1000$ μmol $L^{-1}]$, and very high (VH) $EC_{50} > 1000$ μmol L^{-1}, according to MERCK classification[7]. The advantage of applying uncorrelated principal components of the molecular descriptor sets has resulted in a simple and transparent model.

Conclusions

Applied are chemometric and decision tree models of ILs toxicity based on their molecular descriptors. Toxicity criteria is based on EC_{50} concentrations for inhibition of acetylcholinesterase, In view of very large of molecular descriptors their colinearity was investigated and was found significant average correlation $R^2 \approx 0.4$. In order to simplify and obtain robust models the matrices of cation and anion descriptors are projected to the corresponding spaces of the first ten eigenvectors resulting into about 99.95 of variance (data dispersion content).

Application of chemometric models, partial component regression and partial least squares, resulted in limited quality of prediction on test sets with regression coefficients R^2 of 0.62 and 0.64. However, application of decision tree and random forest models significantly improved quality of prediction with $R^2 = 0.992$. Randomization and aggregation of large population (500 trees) resulted with the model with low overfitting effects and unbiased estimates (besides possible bias in molecule selec-

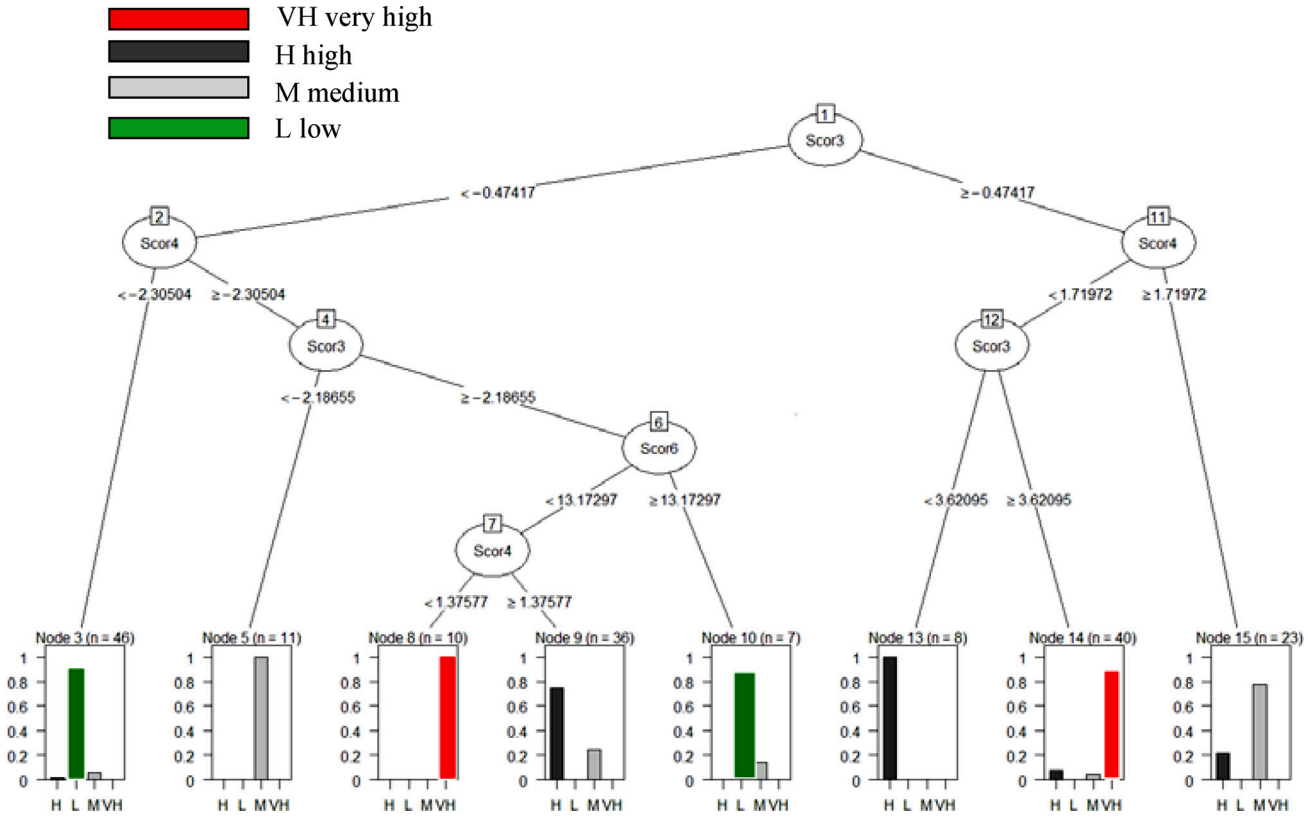

Fig. 4 – *Decision tree model for predictions based on classification of ln(EC_{50}) into VH (very high), H (high), M (medium) and L (low) ILs toxicity categories*

tion). Due to orthogonalisation of training patterns derived are simple and transparent decision trees.

Practical application of the derived models is their potential use as part of a feedback loop for inverse design of new ILs for specific (tailored) new process technology needs.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia, project 058-1252086-0589

List of symbols

A – anions
C – cations
DT – decision tree
E – error matrix
P – vector of principal components by SVD
PCR – principal component regression
PLS – partial least squares
Q – output projection PLS matrix
RFDT – random forest decision trees
SVD – singular value decomposition
T – projection PLS matrix
U – projection PLS matrix
U – input projection PLS matrix
v – eigenvectors
X – matrix of input data
Y – vector of output data
β – vector of model parameters
λ – eigenvalues
σ – standard deviation

References

1. *Kokorin, A.,* Ionic Liquids: Theory, Properties, New Approaches, InTech, Rijeka, Croatia 2011.
2. *Plechkova, N. V., Seddon, K. R., Izgorodina, E. I.,* Theoretical Approaches to Ionic Liquids: From Past History to Future Directions, Ionic Liquids Uncoiled: Critical Expert Overviews, *Plechkova, N. V., Seddon, K. R.* (Eds.), Wiley Online Library (2013). http://dx.doi.org/10.1002/9781118434987
3. SDEWES, 2012, Fraunhofer-Allianz Annual Report, <www.allianz.com/en/investor_relations/results_reports/annual-reports.html>, accessed on 14/12/2013
4. Cvjetko M., Synthesis, application in biotransformations and cytotoxycity of selected imidazolium-based ionic liquids, Doctoral Thesis, (in Croatian) University of Zagreb, Faculty of Food Technology and Biotechnology, Zagreb, Croatia, (2012).
5. *Suojiang, Z., Lu, X., Zhou, Q., Li, X., Zhang, X., Li, S.,* "Ionic Liquids, Physichochemical Properties", Elsevier, Amsterdam, The Netherlands, 2009.

6. NIST, Ionic Liquids Database Standard Reference Database #147, <ilthermo.bouder.-nist.gov /ILThermo/mainmenu.uix>, accessed on 18/07/2013
7. The UFT/ Merck Ionic Liquids Biological Effects Database, <www.il-eco.uft.uni-bremen.de>, accessed on 14/12/2013.
8. *Chun, W. Y.,* J. Comp. Chem. **32** (2010) 1466. http://dx.doi.org/10.1002/jcc.21707
9. *Brereton, G. R.,* J. Chemometrics **28** (2014) 749.
10. *Varmuza, K., Filzmoser, P., "Multivariate Statistical Analysis in Chemometrics",* CRC Press, Baton Rouge, Louisiana, USA, 2009.
11. *Wehrens, R.,* Chemometrics with R, 2011, Springer, New York, USA, 2011. http://dx.doi.org/10.1007/978-3-642-17841-2
12. *Fatemi, M. H., Izadiyan, P.,* Chemosphere **84** (2011) 553. http://dx.doi.org/10.1016/j.chemosphere.2011.04.021
13. *Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P.,* J. Chem. Inf. Comput. Sci. **43** (2003) 1947. http://dx.doi.org/10.1021/ci034160g
14. *Kurtanjek, Ž.,* Proceedings of "24th European Symposium on Computer Aided Process Engineering", *Klemeš, J. J., Varbanov, P. S., Liew, P. Y.,* (Eds.), Budapest, Hungary, 15–19 June (2014) 127. http://dx.doi.org/10.1016/B978-0-444-63456-6.50022-3
15. *Breiman, L., Cutler, A.,*< www.stat.berkeley.edu/~breiman/RandomForests>, accessed on 14/12/2013.
16. R Development Core Team, R: A language and environment for statistical computing. R, Vienna, Austria, <www.R-project.org>, accessed on 14/12/2013.
17. StatSoft, Inc. STATISTICA, v.10. <www.statsoft.com>, accessed on 14/12/2013.
18. *Therneau, T., Atkinson, B., Ripley, B.,* 2013, CRAN – Package rpart, <cran.rproject.org/web/ packages/rpart/index.html>, accessed on 14/12/2013.