ABSTRACT

Berdnyk M. I. L_1-regularization method for the description of the physical and chemical properties of molecules. Qualification scholarly paper: a manuscript.
Thesis submitted for obtaining the Doctor of Philosophy degree in Natural Sciences, Speciality 102 – Chemistry. – V. N. Karazin Kharkiv National University, Ministry of Education and Science of Ukraine, Kharkiv, 2021

This thesis focuses on the study of the possibilities of L_1-regularization application in the construction of "structure-activity" chemometric models and quantum chemical calculations. To perform the tasks of the thesis, an original set of programs has been developed that implement various statistical (chemometric) approaches to the construction of regression models and analysis of their prognostic properties. A set of quantum chemical programs has also been created, in which L_1-regularization is used to construct wave functions of methods that take into account electronic correlation.

In particular, in the thesis we consider application of L_1-regularization to obtain linear empirical models for describing various physicochemical parameters of molecules. Among the such parameters the pKa and boiling points for different nature organic compounds including carbonic acids, phenols, sulfides, fluoroalkanes were investigated. Also the correlations between viscosity of media and vapor pressure for different organic compounds have been described.

Based on the studied samples of molecules, it was shown that with the use of L_1-regularization it is always possible to form a sequential (ordered) set of descriptors. By systematically adding descriptors from this set to linear regression models or artificial neural networks, it is possible to obtain equations (or neural networks, respectively) with successively increasing values of validation criteria. Due to the fact that after ranking of the descriptors set, the selected predictors can be used in different approaches to construct linear regression models, we conducted a corresponding study of the quality of these alternative models. We considered: the *Ordinary Least Squares* method (OLS), the *Least Absolute Deviation* method (LAD), and the *Orthogonal Distances Regression* method (ODR), as well as the recently

proposed method of the *Least Absolute Deviation of Orthogonal Distances* (LADOD). It has been shown that depending on the set of data the different methods can have better prognostic abilities according to the criteria of external or internal validation. It is shown that with the use of artificial neural networks, based on the preliminary ordered by the method of L_1-regularization descriptor set, high-quality predictions of the properties of matter can also be made. The obtained linear regression equations were also compared with alternative approaches that work with non-shrinked (non-optimized) descriptor sets, namely: with the PCR method (Principal Component Regression), as well as with the PLS method (Partial Least Squares or Projection on Latent Structure). It should be noted that although with the use of these methods for some problems we obtained fairly reliable predictive models, however, such models do not provide clear information about the nature of the obtained equations and do not answer the question of what structural and chemical features, or molecular descriptors, lead to changes in response (activity). In the studied examples, we used L_1-regularization to formulate compact one-, two- or three-parametric models that are able to satisfactorily describe the data set. According to the studied examples, the models obtained with pre-selection, using LARS-LASSO, turned out to be better than the results of PLS and PCR calculations.

In the proposed PhD thesis some attention is paid to validation methods and quality of regression equations estimates. For this purpose, a model problem was used in which errors were introduced in both the dependent and independent variables. To simplify the analysis, as well as to study the validation characteristics of the equations obtained in all the studied methods of linear regression, we considered the simplest, but not the trivial case - regression with one independent variable. Such formulation of the problem made it possible to estimate the equations in accordance with the proximity of the coefficients of the regression equations to the "ideal" theoretical values. With the use of the mentioned model problem, the influence of rational sampling on the training and test sets on the quality of the obtained regression equations was investigated. It has been shown that random single sampling is not informative because, depending on the molecules in the test sample, the

validation characteristics for the initial (complete) sample can be both very poor, leading to underestimation, and very good, leading to overestimation of the equation quality. Therefore, it is shown that in order to adequately estimate the quality of the regression equation, as well as to study the quality of the input data in general, it is necessary to create and study as many samplings into a training and test sample as possible. The influence of taking into account the limitations of Applicability Domain (AD) of the model on the validation characteristics of regression equations was also investigated. It is established that, at random divisions of the sample into training and test maxima of distribution of internal and external characteristics, as a rule, coincide. In contrast, when the breakdowns into training and test samples are performed in such a way that the test sample is in the AD of model obtained from the training sample, the corresponding distribution of external validation criteria is shifted relative to the internal in the direction of increase. The most informative are the partitions, which fall close to the maximum density of points. It is from the analysis of these areas one gets the most complete adequate understanding of the quality of the model. The known validation criteria proposed to date were also investigated. Based on the model problem, it was concluded that some of the validation criteria are too highly correlated with each other, which makes their simultaneous use uninformative. Among the following parameters are the pairs: (R^2_{test} - CCC and Q^2_{F3} - $RMSEP_{test}$). It is established that for data with substantial scatter the typical picture is the inverse (essentially nonlinear) dependence $R^2_{train} - R^2_{test}$. In this case, the improvement (increase) of the internal validation coefficients (R^2_{train}, Q^2_{LOO}) is generally not the evidence of an improvement in the oracle properties of the model, because for linear regression at a sufficiently large number of points the relationship between these two criteria was always close to linear. However, the criterion Q^2_{LOO} can be successfully used for small samples. Based on the calculated data, it is shown that for most cases, the OLS method gave the best results. However, for large samples, with errors in both the dependent and independent variables, in the ODR (and LADOD) method, the best equations can be obtained.

Another problem that is closely related to the construction of statistical models is the construction of the classification function. For this purpose, the L_1-regularized calculation of logistic regression was performed in this work. Two problems are considered. In the first classification of molecules on strong and weak bases according to their binding affinity towards Li^+ ion in the gas phase is carried out. In the second problem, organic molecules were classified as active or inactive according to the estrogen receptor relative binding affinity. It is shown that with the use of L_1-regularized logistic regression it is possible to achieve classification results that are competitive with those obtained using other, more complex in the computational sense, methods. The use of a special L_1-regularized algorithm (denoted as LR-LARS-LASSO) made it possible to obtain fairly simple classification equations that are interpretable (in contrast to the results obtained in other popular classification methods such as: Support-Vector Machines, Random Forest, Artificial Neural Networks). Also, the obtained logistic regression equations are unambiguous and reproducible.

It is shown that the L_1-regularization method can be used in quantum chemistry. Using the L_1-regularization procedure, it is possible to create an ordered (ranked) set of electronically excited configurations relative to the Gartree-Fock state. By including a different number of configurations from the created set, it is possible to obtain a progressive set of approximations to the exact calculations of the methods. The method is implemented in the framework of Meller-Plessett's theory of second-order perturbations (MP2) and different levels of the coupled clusters theory. It has been shown that such approximate solutions give fairly accurate values of the energy characteristics of molecules, and the number of configurations in the calculations can be much lower than in calculations using a complete configuration set of the exact method. A number of computational algorithms using first-order multi-step methods have been implemented to effectively solve the corresponding equations of the coupled clusters theory.

Key words: L_1-regularization, QSAR/QSPR, pK_a of organic compounds, boiling point of organic compounds, viscosity of liquids, vapor pressure, Estrogen Receptor

Ligands, Lithium cation basicity, Linear regression, Ordinary Least Squares, Least Absolute Deviations, Orthogonal Distances Regression, Least Absolute Deviation of Orthogonal Distances, Artificial Neural Networks, validation, Logistic Regression, Moller-Plessett's perturbation theory, Coupled Clusters theory.