IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 3, No. 4, December 2014, pp. 150~155
ISSN: 2252-8938

Quantile Regression Neural Networks Based Prediction of Drug Activities

Mohammed E. El-Telbany
Computers and Systems Department Electronics Research Institute, Egypt

Article Info

Article history:

Received Jul 3, 2014
Revised Nov 9, 2014
Accepted Nov 21, 2014

Keyword:

Machine Learning
Prediction
QSAR
Quantile Neural Networks

ABSTRACT

QSAR (quantitative structure-activity relationship) modeling is one of the well developed areas in drug development through computational chemistry. Similar molecules with just a slight variation in their structure can have quit different biological activity. This kind of relationship between molecular structure and change in biological activity is center of focus for QSAR Modeling. Predictions of property and/or activity of interest have the potential to save time, money and minimize the use of expensive experimental designs, such as, for example, animal testing. Intelligent machine learning techniques are important tools for QSAR analysis, as a result, they are integrated into the drug production process. The effective learnable model can reduce the cost of drug design significantly. The quantile estimation via neural network structure technique introduced in this paper is used to predict activity of pyrimidines based on the structure–activity relationship of these compounds which assist for finding potential treatment agents for serious disease. In comparison with statistical quantile regression, the qrnn significantly reduce the prediction error.

Corresponding Author:

Mohammed E. El-Telbany,
Departement of Electrical and Computer Engineering,
Computers and Systems Department Electronics Research Institute, Egypt.
Email: telbany@eri.sci.eg

1. INTRODUCTION

In recent years, with products of human genome project helping to reveal many new disease targets to which drug treatments might be aimed, all the major pharmaceutical companies have invested heavily in the routine ultra-High Throughput Screening (uHTS) of vast numbers of '*drug-like*' molecules guided by chemoinformatics investigations [Lipinski, 2004, Leeson el at., 2004]. The development of a new drug is still a challenging, time-consuming and cost-intensive process and due to the enormous expense of failures of candidate drugs late in their development, uHTS in vitro assays now cover liabilities such as possible side effects [Li, 2005] as well as therapeutic properties, computational methods can be used to assist and speed up the drug design process. It is obvious that the drug discovery and development process would greatly benefit from faster and cheaper procedures to identify chemical compounds with desired biological properties and to optimize their structure in order to obtain effective drugs. Several major bottlenecks in drug discovery may be addressed with computer-assisted drug design methods, such as quantitative structure–activity relationships (QSAR) models [Hansch, 1969], where the molecular activities are critical for drug design, they can be predicted by QSAR models.

Molecular activity is determined by its structure, so structure parameters are extracted by different methods to build QSAR models. Nowadays, machine learning algorithms have been used in the modeling of QSAR problems [Duch et al., 2007, Chin and Chun, 2012; Gertrudesa et al., 2012]. They extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set.

Formally, a learning algorithm is tasked with selecting a hypothesis that best supports the data. Considering the hypothesis to be a function f mapping from the data space X to the response space Y; i.e., $f: X \rightarrow Y$. The learner selects the *best* hypothesis f^* from a space of all possible hypotheses F by minimize errors when predicting value for new data, or if our model includes a cost function over errors, to minimize the total cost of errors.

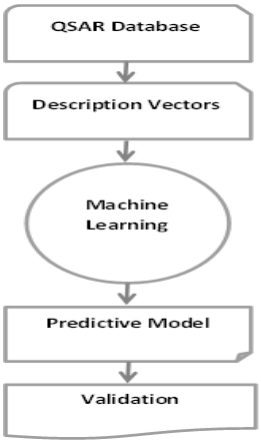

Figure 1. General Steps of Developing QSAR Models.

As shown in Figure 1, the QSAR modeling is heavily dependent on the selection of molecular descriptors; if the association of the descriptors selected to biological property is strong the QSAR model can identify valid relations between molecular features and biological property/activity. Thus, uninformative or redundant molecular descriptors should be removed using some feature selection methods during (*Filters*) or before (*wrappers*) the learning process. Subsequently, for tuning and validation of the predictively of learned QSAR model, one of the validation strategy can be applied likes cross-validation, leave-one-out or the full data set is divided into a training set and a testing set prior to learning (See [El-Telbany, 2014] for a survey). Actually, the machine learning field [Bishop, 2006; Marsland, 2009; Burke and Kendall; 2014, Mohri, 2012] have versatile methods or algorithms such as decision trees, lazy learning, *k*-nearest neighbors, Bayesian methods, Gaussian processes, artificial neural networks, artificial immune systems, particle-swarm optimization, artificial bee optimization, cuckoo search, support vector machines, and kernel algorithms for a variety of tasks in drug design. These methods are alternatives to obtain satisfying models by training on a data set.

However, the prediction from most regression models-be it multiple regression, neural networks, support vector machine (SVM), decision trees, etc. - is a point estimate of the conditional mean of a response (i.e., quantity being predicted), given a set of predictors. Recent advances in computing allow the development of regression models for predicting a given quantile of the conditional distribution, both parametrically and nonparametrically. The general approach is called *Quantile Regression*, but the methodology (of conditional quantile estimation) applies to any statistical model, be it multiple regression, SVM vector machines, or random forests. quantile regression neural networks (QRNN) [Taylor, 2000] estimates conditional values of an individual quantile using a multilayer perceptron neural network. In this paper, quantile regression neural networks techniques are used to study the relationships between different known target activities with respect to the attributes of a finite number of compounds. The A neural network is used to estimate the potentially non-linear quantile models which is more resistance to the outliers. The rest of the paper is organized as follows. Section 2 briefly introduces the QSAR models and quantile regression. Section 3 is introduces the quantile neural networks. Section 4 describes the data set and preprocessing steps. Section 5 describes an evaluation of QSAR modeling and prediction results. Section 6 describes an experimental results. Finally Section 7 presents the findings and conclusions.

2. QSAR MODELS AND QUANTILE REGRESSION

QSAR models are in essence a mathematical function that relates features and descriptors generated from small molecule structures to some experimental determined activity or property [Livingstone, 1995]. The structure-activity study can indicate which features of a given molecule correlate with its activity, thus making it possible to synthesize new and more potent compounds with enhanced biological activities. QSAR analysis

is based on the assumption that the behavior of compounds is correlated to the characteristics of their structure. In general, a QSAR model is represented as follows:

$$activity = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \quad (1)$$

Where the parameters $\boldsymbol{X_i}$ are a set of measured (or computed) properties of the compounds and $\boldsymbol{\beta_0}$through $\boldsymbol{\beta_i}$ are the calculated coefficients of the QSAR model.

Quantile regression was introduced by Koenker and Bassett [1978] as a complement to least squares estimation (LSE) or maximum likelihood estimation (MLE) and leads to far-reaching extensions of "*classical*" regression analysis by estimating families of *conditional quantile surfaces*, which describe the relation between a one-dimensional response $\boldsymbol{y}$ and a high dimensional predictor $\boldsymbol{x}$. Quantile regression offers a number of advantages over least-squares methods. While ordinary least squares regression typically assumes that the error terms are IID, normally distributed, and homoscedastic, quantile regression does not require these restrictive assumptions. Furthermore, since quantile regression estimates quantiles of the conditional distribution rather than the mean, it is more resistant to outliers than least-squares methods [Koenker, 2005]. In contrast to the least-squares loss function $\boldsymbol{L(u) = u^2}$, quantile regression makes use of the asymmetric loss function as shown in Figure 2.

$$\begin{aligned} \boldsymbol{\rho_\tau(u)} &= \boldsymbol{|u| \cdot (\tau \cdot I(u \geq 0) + (1 - \tau) \cdot I(u < 0))} \\ &= u \cdot (\tau - I(u < 0)) \end{aligned} \quad (1)$$

Where $\boldsymbol{\tau}$ corresponds to the quantile to be estimated [Koenker, 2005]. Note that if $\boldsymbol{\tau = 0.5}$, i.e., the median is being estimated, then this loss function becomes simply.

$$\boldsymbol{\rho_\tau(u) = |u|} \quad (2)$$

and the sum of the absolute values of the residuals is minimized to perform regression. To fit a model $\boldsymbol{y_i = \beta^T x_i + \varepsilon_i}$, we estimate $\boldsymbol{\beta}$ using.

$$\boldsymbol{\hat{\beta} = \operatorname{argmin}_{\beta \in R^d} \sum_i \rho_\tau(y_i - \beta^T x_i)} \quad (3)$$

Where $\boldsymbol{d}$ is the number of parameters in our model, so that $\boldsymbol{\beta}$ and $\boldsymbol{x_i}$ are vectors of length $\boldsymbol{d}$. This computation cannot be carried out analytically, in contrast to the computation of least squares regression. Instead, this can be reformulated as a problem in linear programming [Koenker, 2005].

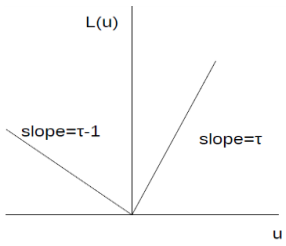

Figure 2. The quantile regression loss function.

3. QUANTILE NEURAL NETWORKS

Artificial neural networks is one of machine learning techniques which have been developed as generalizations of mathematical models of biological nervous systems. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. The learning situations in neural networks may be classified into three distinct sorts. These are supervised learning, unsupervised learning and reinforcement learning [12]. The most widely-used neural network for prediction is

the single hidden layer feed-forward network. It consists of a set of $\boldsymbol{n}$ inputs, which are connected to each of $\boldsymbol{m}$ units in a single hidden layer, which, in turn, are connected to an output (see Figure 3).

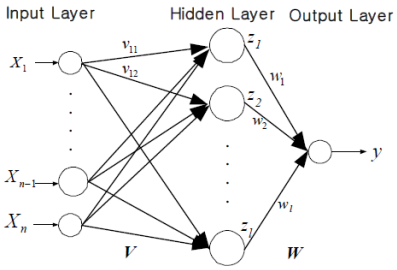

Figure 3. Structure of the neural network.

The resultant model can be written as

$$\boldsymbol{f(x_t, v, w) = g\left(\sum_{j=1}^{m} v_j\, h\left(\sum_{i=1}^{n} w_{ji} x_{it}\right)\right)} \tag{4}$$

Where $g(\cdot)$ and $h(\cdot)$ are activation functions, which are frequently chosen as sigmoidal and linear respectively, and w_{ji} and v_i are the weights (parameters) to be estimated [Taylor 2000].The parameters of the network (i.e. weights) are estimated by optimizing an objective function (e.g., via minimizing least square error). Instead of fitting a linear quantile function using the expression in (5), a quantile regression neural network model, $f(x_t, v, w)$, of the θ^{th} quantile can be estimated using the following minimization.

$$\begin{aligned}\min_{v,w}\Bigg(&\sum_{t|y_t \geq f(x_t,v,w)} \tau \cdot |y_t - f(x_t, v, w)| \\ &+ \sum_{t|y_t < f(x_t,v,w)} (1-\tau) \cdot |y_t - f(x_t, v, w)| \\ &+ \lambda_1 \sum_{j,i} w_{ji}^2 + \lambda_2 \sum_{i} v_i^2\Bigg)\end{aligned} \tag{5}$$

Where λ_1 and λ_2 are regularization parameters which penalise the complexity of the network and thus avoid overfitting [Bishop, 1996; 2006]. The optimal values of where λ_1 and λ_2 and the number, m, of units in the hidden layer can be established by cross-validation [Bishop, 1996; 2006].

4. DATA SET AND PREPROCESSING

The datasets used in this study are obtained from the UCI Data Repository [Newman *et al.*, 1998]. Pyrimidines dataset contains 74 drugs, and each drug has three possible substitution positions. Each substituent is characterized by 9 chemical properties features: polarity, size, flexibility, hydrogen-bond donor, hydrogen-bond acceptor, π donor, π acceptor, polarizability and σ effect. Drug activities are identified by the substituents. The Pyrimidines dataset is randomly shuffled and split into 2 parts in the proportion of 2:1. One part is used as the training set, which contains pairs of 52 compounds. The other part is chosen as the unseen testing set, which contains pairs of the left 22 compounds and those between the 22 compounds and the training 52 compounds. Due to the "*curse of dimensionality*" problem, searching for informative compounds set as a preprocessing step prior to the application of *qrnn* algorithm is important for many reasons. One reason is, that the prediction accuracy of the *qrnn* decreases when irrelevant or radiant features are added. Another problem particularly affecting the computation time is the lacking scalability of the qrnn algorithm. Several approaches to the variable selection problem using information theoretic criteria have been proposed. Many rely on empirical estimates of the mutual information between each variable and the target:

$$+(\boldsymbol{I}) = \int_{x_i}\int_{y} \boldsymbol{p}(x_i, y)\, \boldsymbol{log}\, \frac{\boldsymbol{p}(x_i, y)}{\boldsymbol{p}(x_i)\boldsymbol{p}(y)}\, \boldsymbol{dxdy} \qquad (6)$$

We are selected the first 6 informative compounds.

5. QSAR MODELS VALIDATION

The validation of a qsar relationship is probably the most important step of all. The validation estimates the reliability and accuracy of predictions before the model is put into practice. Poor predictions misguide the direction of drug development and turn downstream efforts meaningless. To verify model quality in regression tasks, predictions are made on the testing set in order to check the agreement between the theoretical values and experimental values by calculating root-mean square error of prediction (RMSE).

$$\boldsymbol{RMSE} = \sqrt[2]{\frac{\sum_{j=1}^{m}(\hat{y}_j - y_j)}{\boldsymbol{m}}} \qquad (7)$$

Where, $\hat{y}_j$, values of the predicted values, and $\boldsymbol{y_j}$, values of the actual values. However, it is necessary to get a large number of testing compounds in order to draw statistically convincing conclusion.

6. EXPERMENTIAL RESULTS

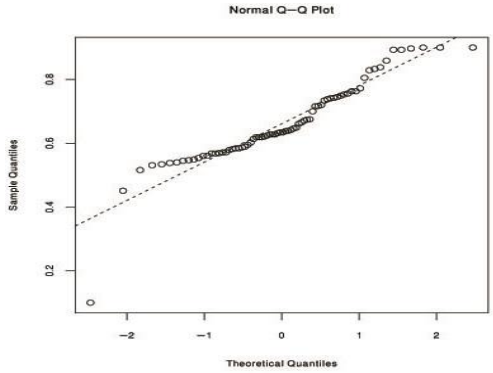

Figure 4. QQ norm probability of drug activity.

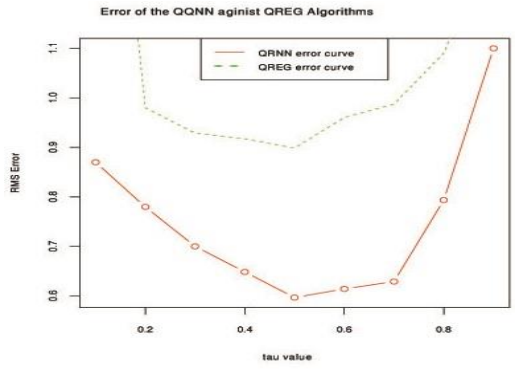

Figure 5. The error curves for distinct quantile using QRNN and QREG.

In this section we present the implementation of QRNN estimates of the effects of drug compound on changes in drug activity. We estimate the model by statistical quantile regression and non-linear QRNN at quantiles 0.1^{th}, 0.2^{th}, 0.3^{th}, 0.4^{th}, 50^{th}, 0.6^{th} , 0.7^{th}, 0.8^{th} and 9^{th}. In the solution process, we used "*quantreg*" and "*qrnn*" packages in R programme implemented linear programming. The QQ normal probability plots is shown in Figure 4, which is used to compare sample data against a theoretical normal distribution. Figure 5 also shows that the corresponding error curves of the QRNN and QREG (for different $\boldsymbol{\tau}$). The minimum RMS error located at $\boldsymbol{\tau = 0.5}$. From the curves, it is clear that QRNN performance is more better than QREG. By training the QRNN model for 200 iteration with number of hidden nodes equal to 5, and comparing the activity of each drug sample with real data (see Fig 6.), the RMSE prediction error was determined be only 0.59. From these results it is concluded that the neural network quantile regression is superior for a complex nonlinear prediction.

7. CONCLUSIONS

The quantile regression is believed to offer a more complete model than the conventional mean regression. Since its birth, quantile regression has been applied to interpret various problems. Quantile regression is popular and comprehensive, but it is rather difficult to apply it directly for forecasting. Hence, this study aims to advance quantile regression for forecasting the drug activities using non-linear neural network model. The results is encourage but there is needs to explore different data sets with large number of samples and explore different feature selection methods.

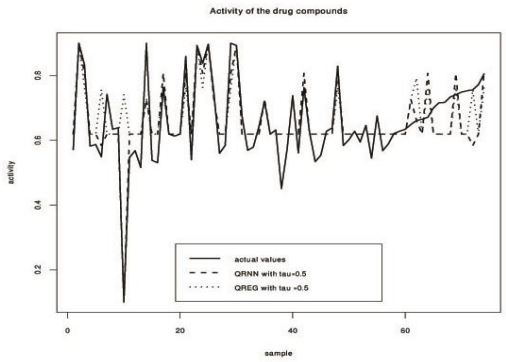

Figure 6. The RMSE of the two method at different quantile.

REFERENCES

[1]. A. Li, Preclinical in vitro screening assays for drug-like properties. *Drug Discovery Today: Technologies*, 2(2):179-185, 2005.

[2]. C. Bishop, Pattern Recognition and Machine Learning, Springer, 2nd, 2006.

[3]. C. Bishop.Neural Networks for Pattern Recognition, *Oxford, Oxford University Press.1996*.

[4]. C. Lipinski, Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies 1(4):337-341, 2004.

[5]. D. Livingstone. Data Analysis for Chemists.-Applications to QSAR and Chemical product Design. *Oxford University Press, 1995.*

[6]. D. Newman, S. Hettich, C. Blake, C. Merz. UCI Repository of Machine Learning Databases, Dept. Information and Computer Science, Univ. California, Irvine, 1998.

[7]. E. Burke and G.Kendall, Search Methodologies Introductory Tutorials in Optimization and Decision Support Techniques, 2nd (ed.) *Springer*, 2014.

[8]. Hansch, A Quantitative Approach to Biochemical Structure-Activity Relationships.*Acct.Chem. Res.* 2: 232-239, 1969.

[9]. J. Devillers. Neural Networks and Drug Design. Academic Press, 1999.

[10]. J. Gertrudesa, V. Maltarollob, R. Silvaa, P. Oliveiraa, K. Honórioa and A. da Silva. Machine Learning Techniques and Drug Design, *Current Medicinal Chemistry*, 19, 4289-4297, 2012.

[11]. J. Hirst, R. King and M. Sternberg. Quantitative Structure-Activity Relationships By Neural Networks and Inductive Logic Programming. I. The Inhibition Of Dihydrofolate Reductase by Pyrimidines. *Journal of Computer-Aided Molecular Design*, vol. 8, no. 4, pp. 405-420, 1994.

[12]. J. Taylor. Quantile Regression Neural Network Approach to Estimating The Conditional Density of Multiperiod Returns. *Journal of Forecasting*, Vol. 19, pp. 299-311, 2000.

[13]. L. Chin Yee and Y. Chun Wei, Current Modeling Methods Used in QSAR/QSPR, in Statistical Modelling of Molecular Descriptors in QSAR/QSPR, 1st,Edited by M. Dehmer, K. Varmuza, and D. Bonchev, Wiley-VCH Verlag GmbH & Co, 2012.

[14]. M. El-Telbany. The Predictive Learning Role in Drug Design. In Journal of Emerging Trends in Computing and Information Sciences, *JETCIS*, Vol. 5, No.3 March 2014.

[15]. M. Mohri, A. Rostamizadeh, and A.Talwalkar, Foundations of Machine Learning, MIT Press, 2012.

[16]. P. Leeson, A. Davis, J. Steele, Drug-like properties: guiding principles for design–or chemical prejudice? Drug Discovery Today: Technologies, 1(3):189-195, 2004.

[17]. R. Burbidge, M. Trotter, B. Buxton and S. Holden. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Computers and Chemistry*, vol. 26, no. 1, pp. 4-15, 2001.

[18]. R. Koenker and J. Gilbert Bassett, Regression quantiles, *Econometrica*, 46, pp. 33–50., 1978.

[19]. R. Koenker, Quantile Regression, *Cambridge University Press*, Cambridge, 2005.

[20]. S. Marsland Machine Learning An Algorithmic Perspective, (Chapman & Hall/CRC, 2009.

[21]. W. Duch, K. Swaminathan and J. Meller, Artificial Intelligence Approaches for Rational Drug Design and Discovery, *Current Pharmaceutical Design*, 13, 2007.