



INSTANCE-BASED REGRESSION MODELS FOR QUANTITATIVE BIOLOGICAL ACTIVITIES USING SUPPORT VECTOR MACHINES AND MULTILINEAR MODELS

Andreas Maunz, Christoph Helma

Center for Data Analysis and Modelling, Albert-Ludwigs-University, Freiburg, Germany



Abstract

Background The automatic mining of similar training instances from diverse data and the construction of robust and interpretable (Q)SAR models is an open major challenge in predictive toxicology. This study compares two different models for the prediction of biological activities of chemicals (e.g. LD50 values) trained on similar training instances using data mining methods. They use a robust instance-based approach with linear fragments as chemical descriptors and activity-specific chemical similarities. A confidence value is associated with each prediction to indicate the applicability domain of the training data.

Methods: A support vector regression model using a novel kernel and a similarity weighted multilinear model with prior objective feature selection and principal components analysis are trained for each prediction on a subset of the training instances similar to the current query structure (neighbors). The applicability domain for each prediction is assessed using chemical similarity and activity values of the neighbors. Three publicly available diverse datasets are validated, two of which have not yet been predicted.

Local Models

This application^a shows that quantitative biological activities of chemicals can be reliably predicted from heterogeneous datasets by mining similar training instances based on activity-specific chemical similarities. It addresses quality criteria such as *robustness* and *applicability domain* in a fully automated fashion.

Global models for a training set often suffer from sparse data and overfitting. Training data often has inconvenient properties, such as

- Non-congeneric chemicals
- Noisy and/or missing activity values
- Skewed (or other non-normal) activity distributions
- Confidential data

Local models derived from a set of *neighbours* yield a distinct model for every single prediction while simple robust approximations at different locations can still approximate a complex function as a whole.

Significant Feature Detection: Linear fragments of unlimited length are used as basic descriptors. The significance of linear fragments with respect to the current endpoint is determined using the *Kolmogorov-Smirnov test*, assigning *p*-values to every fragment.

Activity-specific Similarity: Chemical similarity between two compounds \mathbf{x}_i and \mathbf{x}_j is determined by the *p*-weighted fraction of significant features f with $p_f > 0.9$ shared with the query structure.

$$sim(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{f \in F} \{p_f | f \subseteq \mathbf{x}_i \wedge f \subseteq \mathbf{x}_j\}}{\sum_{f \in F} \{p_f | f \subseteq \mathbf{x}_i \vee f \subseteq \mathbf{x}_j\}} \quad (1)$$

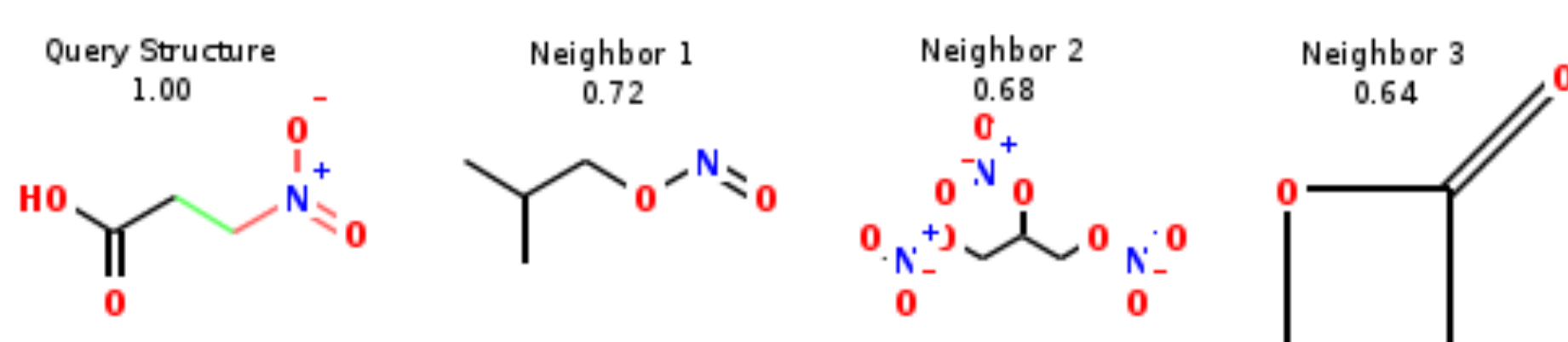


FIGURE 1: Query structure and three neighbours with similarity values.

^aLazar is available from the subversion repository at svn://www.in-silico.de/lazar/trunk

For model training, only *neighbours* \mathbf{x}_i to the query structure \mathbf{x}_q are considered (endpoint-specific similarity $sim(\mathbf{x}_i, \mathbf{x}_q) > 0.3$). Two models were developed for regression:

WEIGHTED MULTILINEAR MODEL

Obtain data matrix X and activity vector y , where y_i denotes activity of training compound i and X_{ij} indicates presence or absence of fragment f_j in neighbour x_i .

Feature Generation and Selection: Transform data by objective feature selection and principal components analysis:

- Exclude features that objectively carry no specific information (ordered by *p*-values)
- Projection-based dimensionality reduction to further reduce noise

Similarity Weighting: The allowed error of the fit for every neighbour in the multilinear model ($\mathbf{y} = \langle \mathbf{b}, \mathbf{X} \rangle + d$) is inverse proportional to its chemical similarity to the query structure $sim(\mathbf{x}_i, \mathbf{x}_q)$.

KERNEL MODEL

Support vector regression (ν -svr with $\nu = 0.8$, see Schölkopf and Smola, 2002) using the activity-specific similarity as kernel function. The gram matrix $\mathbf{K} := \mathbf{K}_{ij} = sim(\mathbf{x}_i, \mathbf{x}_j)$ is used for training and the result is a sparse collection of neighbours and an associated weight vector α . The predictive equation is given by

$$f(\mathbf{x}_q) = \sum_{i=1}^n \alpha_i y_i sim(\mathbf{x}_q, \mathbf{x}_i) + d, \quad (2)$$

Applicability Domain

Applicability Domain Estimation: Every (Q)SAR model has only a limited domain of applicability, namely “the physico-chemical, structural or biological space on which it has been trained” (Jaworska et al. 2003). In this implementation, the estimation is based on two variables, namely \tilde{s} , the median similarity of the neighbours, and σ_a , the standard deviation of the neighbour’s activities.

$$conf \propto \frac{\tilde{s}}{\sigma_a}. \quad (3)$$

Confidence: To actually estimate the membership to the applicability domain, a confidence is implemented as

$$conf = \tilde{s} e^{-\sigma_a}, \quad (4)$$

which fulfills the above requirements. For the multilinear model, the median mahalanobis distance is also reported.

Experiments

Three diverse data sets^a were used to assess the predictivity of the models: *Fathead Minnow* acute toxicity (EPAFHM, 573 compounds), maximum recommended therapeutic dose (FDAMDD, 1215 compounds, **previously unpredicted**), human health effects (IRIS dataset, 68 compounds, **previously unpredicted**).

- **Weighted accuracy (wa):** fraction of correct predictions with respect to the 1 log unit error margin (see Benigni et al. 2007).
- **Q² (q-sq):** squared correlation coefficient between predicted and database activities.
- **Mean error (me):** mean of the raw prediction errors (i.e. not standardized).

^a<http://www.epa.gov/acct/dsstox/>

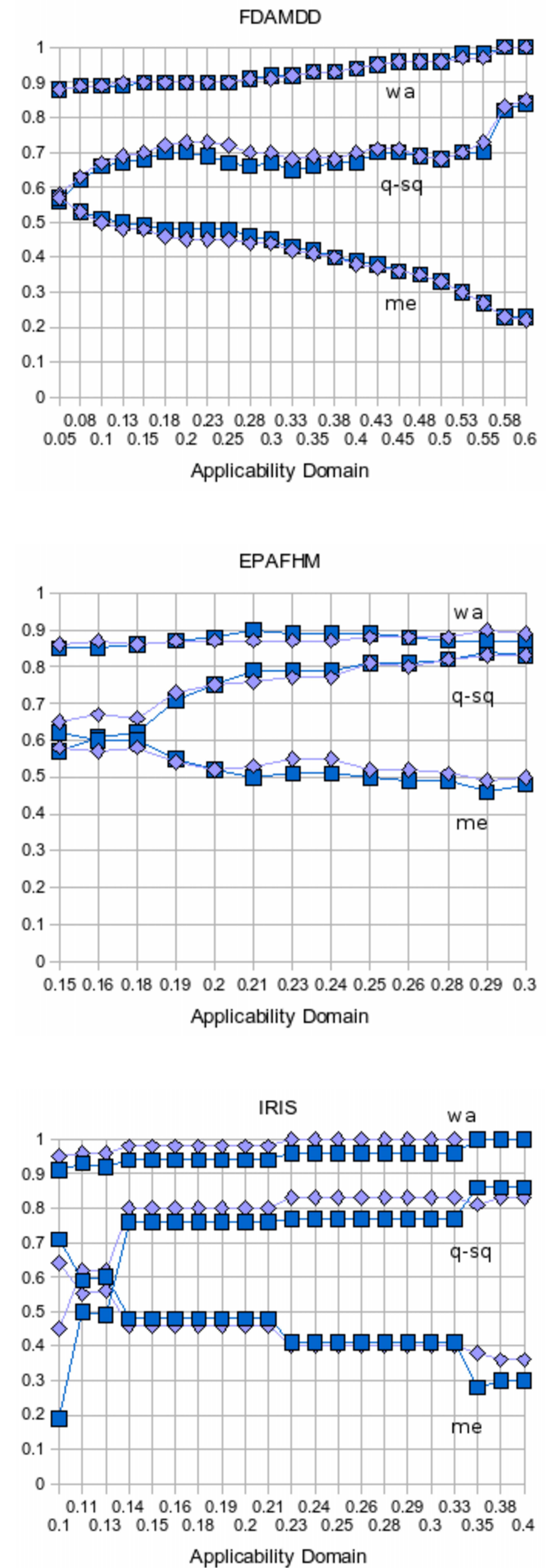


FIGURE 2: Leave-one-out crossvalidation performance using different AD thresholds for the three datasets. Dark blue (lying squares): kernel model, light blue (squares standing on corners): multilinear model.

Discussion: The results indicate a good estimation of prediction quality and also justify the application of the multilinear model (yields results comparable to kernel model with much larger predictive capacity). The instance-based methodology reduces the risk of overfitting and sparse data inherently present in global models.

Use: Potential use includes the filling of data gaps, validation of experimental data, screening, ranking and priority setting and highlighting chemicals of concern (even before their synthesis).