University of Tartu

Introduction to Data Science (LTAT.02.002)

Report for homework 10

Aleksandra Panfilova

Alina Rekena

Prasoon Kumar

Bioconcentration factor: predicting environmental risks of chemicals using classification models

Task 2. Business understanding

(794 words)

Scientists from the University of Tartu (UT) in their earlier research[1] have developed several quantitative structure-activity relationship (QSAR) models using machine learning approaches to predict the bioaccumulation potential of chemicals. Although the accuracy of these models is shown to be at 87% maximum, no other predictive learning algorithms except Random Forest have been tried within these QSAR models. In the field of QSAR, the use of predictive machine learning algorithms is gaining popularity because they provide for the opportunity of automatic and fast prediction of bioaccumulation potential of chemicals. According to REACH legislation, for chemicals produced or imported in the European in amounts greater than 100 tonnes in a year, bioaccumulation measurements are required. The Regulation encourages the use of QSAR models as part of an integrated testing strategy as alternatives to animal tests.

Firstly, project goal is to develop more effective prediction capability of QSAR model. As a basis of evaluation, reference Piir et al., 2014 is used. Secondly, we aim to get more data on the correlation between chemical structure parameters and bioaccumulation.

As our main goal is to improve overall model performance, increase in accuracy of the model (in numeric terms) is the criterion for success. We would consider our second goal

[1] "Classifying bio-concentration factor with random forest ... - NCBI." 6 Dec. 2014, https://www.ncbi.nlm.nih.gov/pubmed/25482723. Accessed 28 Nov. 2019.

achieved if we could find and explain the connection between 3 features found in the original project and the ones we selected.

The dataset is partly publically available (stored in the public repository QsarDB.org by Piir et al., 2014). But as we aim to improve the model, we have acquired from the authors the dataset of full QSAR model (is in our storage). It was needed because the public dataset contains only few features selected by RF algorithm.

For this project we will use the free Jupyter Notebook software with its libraries.

As we have the data already in our storage, the only requirement we have to fulfill is finishing the project by the deadline set within the course "Introduction to Data Science". As we are using different software (Python) and various computing environments (Mac and Windows) than in the reference publication, for the acceptance of our results we are going to refer to a "standard deviation" derived from a replicate model using RF algorithm. We have no other specific obligations regarding the protection or publication of the acquired result, except that we were kindly asked to present our result to the authors of reference publication.

The only risk could be that our personal computers could not have enough computing power to do some computation-intensive models, such as, Support Vector Machine (SVM) with polynomial kernels etc. In that case, we plan to contact the authors of the reference publication and ask for help in terms of the resources.

Glossary:

QSAR model - a mathematical model often used in chemical and biological sciences to help predict biological fate properties of chemical compounds from the knowledge of their chemical structure.

Bioaccumulation - process where chemicals from the environment accumulate in an organism.

Bioconcentration - laboratory equivalent for bioaccumulation process.

Bioconcentration factor (BCF) - ratio of the concentration of a chemical in an organism and that in the surrounding environment at steady state widely used to assess the environmental risks of chemicals (experimental data).

REACH regulation - Registration, Evaluation, Authorisation and Restriction of Chemicals is a European Union regulation dating from December, 2006, which addresses the production and use of chemical substances and their potential impact on both human health and the environment.

Molecular descriptors - is the way molecules are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule. This information can be experimental and theoretical, such as polarizability, dipole moment, etc.).

As this project is a part of university course, costs and benefits are not applicable.

Data-mining:

One of our practical goals is to re-select the 3 most influential features according to SVM coefficients (feature importance). Before feeding data to the model, correlation between features of the data set will be found and only one feature from tightly correlated group will stay in the dataset.

Another practical goal is to try several different machine learning algorithms which are used for QSAR (Support Vector Machine, Naive Bayes) with different parameters and compare performance of resulted models in classification of compounds as bioaccumulative in validation set.

The success criteria for feature re-selection is either getting the same 3 most influential features as Piir et al., 2014 or something tightly correlated to them, as feature selection algorithm of the Sklearn library might give the different result in comparison to the R used in the original work.

The success criteria for the predictor is achieving better performance metrics of the model than the existing ones. As false-negative classification of a compound is potentially more dangerous, increase sensitivity is more preferable then in specificity.

Task 3. Data understanding

(676 words)

QSAR classification model requires the dataset with instances, features and 2 classes. The dataset was provided by the scientists of UT in *csv* format and is already in our storage. Reading the data with Jupyter notebook was successful. Later, when we have the first results, we will use the QsarDB.org database where the performance metrics of all models by Piir et al., 2014 are stored to compare our results. The QsarDB repository provides analytical results for three models developed in the reference publication in the form of three confusion matrices per model.

In QSAR modelling, the predictors (X) and response variables (Y) can mean several different things. In our case, the predictors are theoretical molecular descriptors of chemicals. The response variable could be some biological activity of chemicals. In our case, it is bioaccumulation potential expressed in BCF (logarithmic scale). It is an experimental value. In their publication, UT scientists reveal from were the dataset was obtained: “The data set originally is adapted from the literature[2], and contains 1036 compounds where experimental values are measured inside the pH limits defined by the guidelines in the REACH legislation”. In principle, predictor variables of the QSAR model are created using some other advanced computational approaches that were not part of our project. We did not have to do this computational task. All predictors were provided for us in the dataset.

Instances: 1007 chemical compounds, their names being translated in SMILES notation. (Authors of the reference publication cleaned the dataset, as a result of which the number of instances decreased from 1036 to 1007.)

Features: 551 molecular descriptors (predictors) being columns of the dataset, except column for response variable log(BCF).

Classes: 2 labels (response variable) derived from log(BCF) column. The threshold above which the compound is classified as bioaccumulative is set by the legislation[3], and taking into consideration errors in BCF measurements, is interpreted as log(BCF) = 3.0.

We have been able to convert log(BCF) to classes successfully, and data corresponds to the reference publication.

Instances have been denoted in ID numbers, so we do not see in the dataset chemicals behind them. The cleaning of data to remove chemicals that do not lend themselves for modelling has been done by authors of the reference publication. As a result of their work, chemical mixtures, duplicates, tin and silicon containing compounds and others have been already discarded (in detailed described in the reference publication).

Features are numerical characterization of chemical structures of the compounds (like number of hydrogen, carbon, nitrogen atoms), chemical properties of the compounds (such as mass, ionization potential, electronegativity, charge, number of double, triple and quadruple bonds) and molecular properties of the compounds (such as molecular distance between the carbon, nitrogen and oxygen atoms, Welner Polarity number, Gravitational index of the atoms). The value of the features vary depending on the physico-chemical properties of the compounds. One example could be the number of hydrogen atoms of the compounds. There are at least 7 compounds with more than 50 hydrogen atoms and there are at least 18 compounds which does

[2] "A new bioconcentration factor model based on SMILES ... - NCBI." June 17. 2010, https://www.ncbi.nlm.nih.gov/pubmed/20599297. Accessed Dec 1. 2019.

[3] "PBT/vPvB assessment - ECHA - Europa EU." https://echa.europa.eu/documents/10162/13632/information_requirements_r11_en.pdf. Accessed 28 Nov. 2019.

not have any hydrogen atoms. The other example could be the Weiner Polarity Number (WPATH) feature. The values of this feature range from 1 to 30600.

There are many features with 0 value which means that our plan of feature correlation should work, and many of the features will be sorted out and grouped as having similar correlations. The dataset is clean from "NaN" values.

Classes have been labeled as (0), which means non-bioaccumulative or "nB", if log(BCF)<3.0 and (1), which means bioaccumulative or "B", if log(BCF)>3.0. Altogether, 798 (nB) and 209 (B) compounds have been identified.

The distribution of labels for classes indicates that we have an imbalanced dataset, and the majority of the compounds represent the non-bioaccumulative (nB) class. This is going to be addressed in our work the same way as in the reference publication, and we are going to work on 3 different models - imbalanced model towards "nB", balanced model, and imbalanced model towards "B".

Task 4. Planning your project

1. Meet with Uko Maran and Sulev Sild, co-authors of Piir et al., 2014. (1 h, Aleksandra and Alina)

   Output: the main idea (improvement of their models), the discussing dataset used in the paper, suggestion to try SVM.

   Deadline: done.

2. Replicating Piir et al., 2014 models using Pandas, Numpy and Sklearn Python libraries. (8 h, all team members)

   2.1. Read the data and leaving only 3 features used by reference publication

   2.2. Train-test split: sorting samples based on log(BCF) value and taking out each third sample to validation (test) set.

   2.3. Make the output binary by mapping all samples with log(BCF) above 3.0 as bioaccumulative (0) and all below as non-bioaccumulative (1).

   2.4. Creating a balanced dataset by undersampling nB samples.

   2.5. Creating a dataset, inbalanced towards B in 4:1 ratio.

   2.6. Train 3 Random Forest algorithms on imbalanced towards nB, balanced and imbalanced towards B training sets.

   2.7. Predict validation set outputs with all 3 models and comparing confusion matrices to the reference results.

2.8. Calculate difference between our models' performance and reference paper.

Output: "standard deviation" which comes from different programming language usage. It will be used further to estimate the success of the project.

Deadline: done til 2.5 h. Deadline for week Dec 2-8.

3. Re-selecting 3 most important features (Aleksandra, 4 h)

3.1. Find correlations between the features and reducing the feature number by getting rid of highly correlated features (leaving only 1-2 features).

3.2. Run 100 SVM models on randomly split datasets and check absolutes of coefficients. Look for features which occur among 20 most influential features in at least 66 cases.

Output: Selected features to work further with.

Deadline: Dec 2-8.

4. Optimization of SVM and Naive Bayes models trying different parameters (all team members, Aleksandra 13 h, Alina 17 h, Prasoon 22 h)

4.1. Find out which kernel and other parameters gives the best result by testing different sets of those parameters.

Output: The final model with the best prediction.

Deadline: End of project.

5. Preparing the poster for presentation (4 hours all team members)

5.1. Making good visualisation of research problem, e.g. examples of chemicals that are bioaccumulative, representing how to translate them in dataset (SMILES notation)

5.2. Making description about numeric conversion of molecular descriptors

5.3. Poster design