Construction of ML Models for Toxicity Prediction on Nuclear Receptor AhR

Felix Moorhoff, Iliana Corona Viramontes, Mehrshad Roshanaei & Shrabanti Nath

Technical University of Munich – TUM Campus Straubing

Introduction

AhR is a nuclear receptor member of the bHLH family of transcript factors that regulate physiological and development processes and that can be activated by exogenous chemical compounds that may further lead to adverse responses in the human body, Sorg [1]. These exogenous chemicals are cataloged as toxic and the exposure of these molecules to the environment and human beings is nowadays a global concern.

Compound toxicity is usually measured by in vivo assays, as currently applied for AhR receptor, nevertheless in silico methods represent a low time and cost alternative. In this poster, a predicting model for the binary classification of chemical compounds into toxic and non-toxic will be reviewed based on their chemical and physicochemical descriptors for the aryl hydrocarbon receptor (AhR).

Data Structure

For the design, training and testing of the model three different datasets as csv files were used as:

- First dataset contained 12,000 chemical compounds each with 800 chemical and physicochemical descriptors, that will be called features along this poster, was used for model training.
- Second dataset provided the same 12,000 chemical compounds but considering 12 different receptor responses. For the matter of this work it was only consider the nuclear receptor aryl hydrocarbon receptor (NR. AhR) as toxic effect response.
- Third dataset shared the same structure as the first dataset, considering as the only differences the chemical compound name along with their feature results. This data collection was later used for testing in the final predicting model.

Data Preprocessing and behavior

- Sample imputing
  - All chemical compounds presenting missing labels were removed from the first and second dataset prior their use on the training model.
- Normalizing data by Zero Mean Unit Variance
  - Data standardization was executed in the training data to be then applied on the testing dataset.

$$\text{ZScore} = \sqrt{\frac{x-\mu}{\sigma}}$$

Data Overview

- Visualization of data compressing by Kernel Principal Component Analysis (PCA)
  - Considering a variance explained of α=95% data of first dataset was compressed as:

$$x \in \mathbb{R}^{12060x801} \rightarrow \tilde{x}_2 \in \mathbb{R}^{12060x80}$$

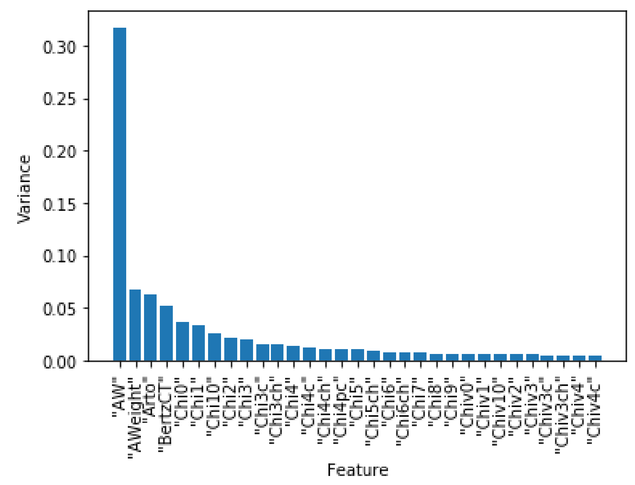

Chart 1. Variance explained by different features on the first dataset, where PC_1 accounts for 31.73% and PC_2 for 6.66% of data.

- Data plotting by toxic, nontoxic and unknown chemical compounds based on second dataset.

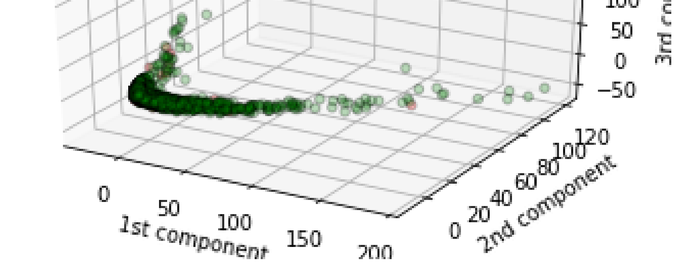

Figure 1. Graphic visualization of data classification into toxic (red), nontoxic (green), and unknown (blue) from second dataset.

- Visualization of a non-linear reduction by Kernel Principal Component Analysis (KPCA)
  - Based on a 3 component Kernel matrix the distribution of toxic and nontoxic chemical compounds from the second dataset was plotted, see Figure 2.

Figure 2. 3-Dimensional plot of a three component linear Kernel for the classification of toxic (red) and nontoxic (green) effects on the chemical compounds of second dataset.

Results

Based on Z-Score normalized and PCA compressed data, a SVC model was computed utilizing a nested-cross validation (CV).

KNN and LR models for comparison were also computed on the same basis but employing KPCA, see Table 1.

Table 1. Results comparison between Logistic Regression (LR), Support Vector Machine (SVC) and K-Nearest Neighbor (KNN) models on training/testing datasets.

Model	MCC	ACC	PPV	TPR	AUC
SVC	0.59	0.92	0.71	0.56	0.77
KNN	0.49	0.90	0.62	0.44	0.72
LR	0.44	0.90	0.66	0.36	0.70

Considering SVC model predictions as the most adequate it was then evaluated on an additional dataset with hidden labels at a Web Server, see results comparison between datasets in Table 2.

Table 2. Comparison of Matthew's Correlation Coefficient (MCC), Accuracy (ACC), Precision (PPV), Recall (TPR) and Area under the curve (AUC) results for the Support Vector Machine on training/testing dataset and Leaderboard dataset.

	Training and Testing Dataset	Leaderboard Dataset
MCC	0.59	0.51
ACC	0.92	0.89
PPV	0.71	0.53
TPR	0.56	0.62
AUC	0.77	NA

Discussion

Even though PCA contemplates fewer number of features after data compressing compared to KPCA, Musa [2], it provided a good basis for an accurate label classification in SVC model.

Despite presenting overfitting on the Leaderboard Dataset, PCA with SVC Model provided the best results overall, differing only by 9% from AUC values obtained by Drwal [3].

Conclusion

The SVC model suited predictions for toxicity labeling in chemical compounds based on their chemical and physicochemical descriptors for the NR. AhR.

References

1. Sorg, O. (2014). AhR signalling and dioxin toxicity. Toxicology Letters, 230(2), 225–233. doi:10.1016/j.toxlet.2013.10.039
2. Musa, A. B. (2013). A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression. International Journal of Machine Learning and Cybernetics, 5(6), 861–873. doi:10.1007/s13042-013-0171-7
3. Drwal, M. N., Siramshetty, V. B., Banerjee, P., Goede, A., Preissner, R., & Dunkel, M. (2015). Molecular similarity-based predictions of the Tox21 screening outcome. Frontiers in Environmental Science, 3. doi:10.3389/fenvs.2015.00054