QSAR Biodegradation

ANJUMANARA ATHINA (155142), NICOLAS CHUANG HUI JIANG (152504), LIM WEI SEN (154151), ONG JIA JIE (153818)

Many regulatory laws resulting from the enactment of the United Nations Stockholm Convention in May 2004, together with the new REACH legislation, have promoted significant new activity in the assessment of Persistent, Bio accumulative and Toxic (PBT) substances. These are chemicals that have the potential to persist in the environment, accumulate within the tissues of living organisms and, in the case of chemicals categorized as PBTs, show adverse effects following long-term exposure. Under REACH, estimated data generated by (Q)SARs may be used both as a substitute for experimental data, and as a supplement to experimental data in weight-of-evidence approaches.

Aim : To use QSAR data from chemical compounds and build QSAR models to predict ready biodegradation of chemicals using different modeling methods and type of molecular descriptors.

Dataset Description

The QSAR biodegradation dataset was built in the Milano Chemometrics and QSAR Research Group (UniversitÃ degli Studi Milano â€“ Bicocca, Milano, Italy). The research leading to these results has received funding from the European Communityâ€™s Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project. . Biodegradation experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE).

	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	...	C-026	F02[C-N]	nHDon	SpMax_B(m)	Psi_i_A	nN	SM6_B(m)	nArCOOR	nX
0	3.919	2.6909	0	0	0	0	0	31.4	2	0	...	0	0	0	2.949	1.591	0	7.253	0	0
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	...	0	0	0	3.315	1.967	0	7.257	0	0
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	...	0	0	1	3.076	2.417	0	7.601	0	0
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	...	0	0	1	3.046	5.000	0	6.690	0	0
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	...	0	0	0	3.351	2.405	0	8.003	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1050	5.431	2.8955	0	0	0	2	0	32.1	4	1	...	0	6	1	3.573	2.242	1	8.088	0	0
1051	5.287	3.3732	0	0	9	0	0	35.3	0	9	...	0	3	0	3.787	3.083	3	9.278	0	0
1052	4.869	1.7670	0	1	9	0	5	44.4	0	4	...	4	13	0	3.848	2.576	5	9.537	1	0
1053	5.158	1.6914	2	0	36	0	9	56.1	0	0	...	1	16	0	5.808	2.055	8	11.055	0	1
1054	5.076	2.6588	2	0	0	0	4	54.5	0	0	...	2	0	0	4.009	2.206	0	9.130	0	2

Table 1: Samples of the dataset

Data Analysis

The dataset is analyzed in order to gain insight from the dataset. Data visualization technique such as scatter plot is used to analyze the relationship between two features. Examples of the plotting are given in Figure 1.

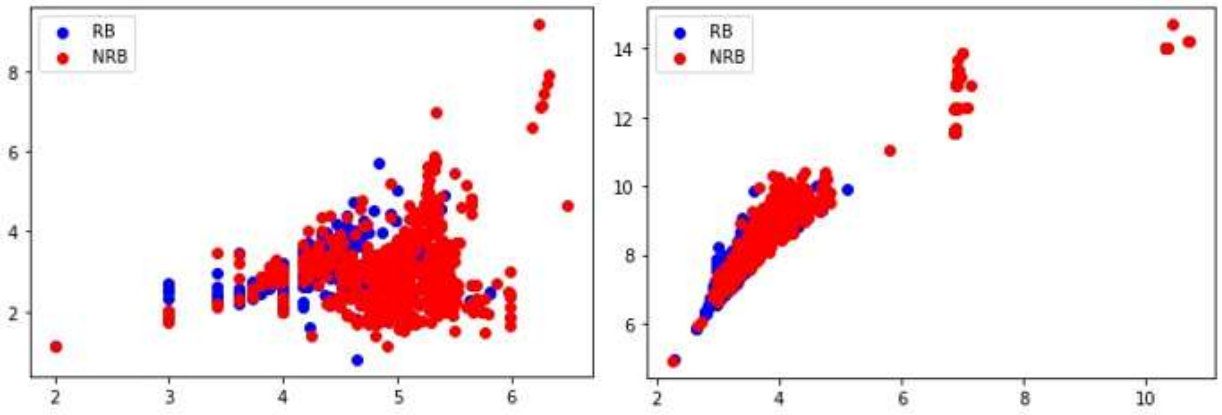

Figure 1: (Left) SpMax_L against J_Dz(e), (Right) SpMax_B(m) against SM6_B(m).

The bivariate analysis is performed to determine the relationship between the features and target. The bivariate analysis is performed by performing the feature selection using ANOVA. ANOVA is chosen since the features are numerical data and the target is categorical data. Six features are selected out of twenty features. Table 2 shows the selected features.

Feature	Type	Value/Statistics
SpMax_L	Numerical continuous	Range: 2 - 6.496 Mean: 4.7826 Std: 0.5469
J_Dz(e)	Numerical continuous	Range: 0.8039 - 9.1775 Mean: 3.0695 Std: 0.8313
SpMax_B(m)	Numerical continuous	Range: 2.267 - 10.695 Mean: 3.9182 Std: 0.9996
SM6_B(m)	Numerical Discrete	Range: 4.917 - 14.70 Mean: 8.629 Std: 1.2419
nHM	Numerical Discrete	Range: 0 - 12 Mean: 0.7166 Std: 1.4624
nN	Numerical Discrete	Range: 0 - 8 Mean: 0.6862 Std: 1.09038

Table 2: The selected features

Data Modeling

Three predictive models are built using Decision Tree, SVM and Neural Network algorithms. The models are evaluated using hold-out method. The ratio of the split is 70% training set and 30% test set. The parameters of the predictive models are given in Table 3.

Algorithm	Value/Statistics
Decision Tree	Criteria: Gini Max Depth: 7 Min Samples in Leaf: 1
K-Nearest Neighbor	K: 16

Table 3: Parameters of the predictive models.

The result of the classification of each predictive models are given below.

```
0.8388625592417062
[[133  19]
 [ 15  44]]
              precision    recall  f1-score   support

         NRB       0.90      0.88      0.89       152
          RB       0.70      0.75      0.72        59

    accuracy                           0.84       211
   macro avg       0.80      0.81      0.80       211
weighted avg       0.84      0.84      0.84       211
```

Figure 2: Results of classification using Decision Tree model.

```
0.8636363636363636
[[162  21]
 [ 15  66]]
              precision    recall  f1-score   support

         NRB       0.92      0.89      0.90       183
          RB       0.76      0.81      0.79        81

    accuracy                           0.86       264
   macro avg       0.84      0.85      0.84       264
weighted avg       0.87      0.86      0.86       264
```

Figure 3: Results of classification using K-Nearest Neighbor model.

For classification, we can do this by calculating the correct and incorrect classification and put them in a table called confusion matrix. The total number of correctly identified in yes class is true positive while the opposite is true negative. The number of instance which

belongs to yes but misclassified as no is false negative and the opposite is false positive.

	Yes	No
Yes	True Positive (TP)	False Negative (FN)
No	False Positive (FP)	True Negative (TN)

The formula to calculate precision is shown below:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

For Decision tree, the precision for Not-ready biodegradable(NRB) is 0.90, while the precision for Ready biodegradable (RB) is 0.70. For the macro avg, it is 0.80 and for the weighted avg, it is 0.84.

For K-Nearest Neighbor, the precision for Not-ready biodegradable(NRB) is 0.92, while the precision for Ready biodegradable (RB) is 0.76. . For the macro avg, it is 0.84 and for the weighted avg, it is 0.87.

These results show that using K-Nearest Neighbor algorithms have a higher precision than the decision tree which means using K-Nearest Neighbor is more accurate or precise as many of them are actual positive. In other words, for QSAR biodegradation, a false positive means that NRB has been identified as RB for RB case, this shows that the result is not precise when the precision is not high.

The formula to calculate recall is shown below:

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

For Decision tree, the recall for Not-ready biodegradable(NRB) is 0.88, while the recall for Ready biodegradable (RB) is 0.75. For the macro avg, it is 0.81 and for the weighted avg, it is 0.84.

For K-Nearest Neighbor, the recall for Not-ready biodegradable(NRB) is 0.89, while the recall for Ready biodegradable (RB) is 0.81. For the macro avg, it is 0.85 and for the weighted avg, it is 0.86.

Recall calculates how many true positives are there in total for actual positive. The results show that using K-Nearest Neighbor model have a high value of recall than the decision tree model.

Based on these results, we notice that K-Nearest Neighbor algorithm has higher precision and higher recall than Decision tree algorithm. Therefore, we can say that K-Nearest Neighbor algorithm is a better algorithm than decision tree.

However, we cannot just determine which algorithm is the best based on precision and recall only. We must also look into other perspectives such as accuracy and f1-score. F1-score is other measure which is used to see the balance or equilibrium between the precision and recall. Therefore, it is important for us to consider all of this to decide which is the best algorithm.