

Development of the KY-methods for use on toxicity prediction O Kohtaro Yuta In Silico Data, Ltd. (Japan) (http://www.insilicodata.com)

Research object : Developing new powerful data analysis methods that are specialized in toxicity evaluation

Main difficulty to execute toxicity evaluation 1. Structural diversity of compounds is quite large 2. Number of samples used in the analysis is very large 3. Quite complex of the toxicity expression mechanism 4. High classification and prediction value is required

Data analysis technique that is normally used is a shortage in toxicity evaluation

Suitable for toxicity prediction

Data analysis techniques that is normally used are shortage in toxicity evaluation

New data analysis methods (KY-methods) have been developed in order to solve toxicity evaluation problems

Outstanding features of the KY(K step Yard sampling)-methods

Binary classification

- 1. Constantly achieve perfect (100%) classification under any conditions
 - Highly overlapped class sample data set
- •Quite large number of sample data set (tens and several thousands of)

Repeat these operation, until

Spatial region on sample space

all samples are correctly classified

- 2. Starting sample set was divided into
 - small and clean sample set
 - small and hierarchical sample set

Sample space : highly overlapped space



Fitting : Regression analysis 1. Constantly achieve high coefficient of correlation and high decision coefficient under any conditions •Widely distributed sample data space •Quite large number of sample data set (tens and several thousands of) 2. Starting sample set was divided into •'inlier' and 'outlier' sample set **Repeat these calculation, until** small and hierarchical sample set no more can this operation It is very difficult to achieve a high coefficient of determination and Starting sample space correlation coefficient by this sample spa

1. Two model KY- discriminant method 2. One model KY- discriminant method 3. Model free KY- discriminant method

2. Three zone KY-fitting method

List of application examples of the KY method and conclusions

Example 1 (Binary data) : Ames test sample data set About 7000 Ames test sample dataset \Rightarrow Perfect classification * Usual multi-variate and pattern recognition methods can't

Example 2 (Binary data) : Skin sensitization sample data set About 600 sample dataset \Rightarrow Perfect classification * Usual multi-variate and pattern recognition methods can't * Poster P05-21 Sato et.al, Euro Tox 2013

* All six KY-methods are patented and pending applications in JP, USA and EU and some were Korea and China.

Example 3 (Continuous data): Fish toxicity sample data set About 800 Fish toxicity sample dataset ⇒ High decision coefficient * Usual multi-variate and pattern recognition methods can't

Results and conclusions:

Very high correlation coefficient and perfect classification (100%) were realized by newly developed KY-methods. On the toxicity research field, it is almost impossible and hard to achieve excellent and stable evaluation results by using the conventional data analysis methods.

All Rights Reserved, Copyright © INSILICODATA LIMITED 2013



Big toxicity data analysis by the combination of the PCA method and the KY-methods

In Silico Data, Ltd. (http://www.insilicodata.com) **O Kohtaro Yuta**

• Apply multi-step and re-sampling technologies by the KY-methods to the PCA for handling big toxicity data •

🛞 ADMEWORKS ModelBuilder - [fukui2010-593all 174neg 419pos step1 AN and APmodels only.amb:2]	🛛 💽 ADMEWORKS ModelBuilder - [fukui2010-593all 174neg 419pos step1 AN and APmodels only.amb:2] 📃 🗖 🔀
🕑 <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>S</u> ample <u>D</u> escriptors F <u>e</u> atureSelection <u>G</u> raphs <u>M</u> odel <u>W</u> indow <u>H</u> elp	🗹 🕑 <u>F</u> ile <u>E</u> dit <u>V</u> iew <u>Sample</u> <u>D</u> escriptors F <u>e</u> atureSelection <u>G</u> raphs <u>M</u> odel <u>W</u> indow <u>H</u> elp _ <u>_</u> 🗗 🗙
Model name: Skin_CL_ADA_1 Sample Dot Map Zoom	PC Scatter PC Vectors
Number of training : 50 Percent features in descriptor set : 75.00 Percent correct sample : 82.125 Number of samples : 593 Number of models : 47 TRAINING SET CLASSIFICATION Class Members Wrong %Correct neg 174 22 87.36 pos 419 84 79.95 Total 593 106 82.12	20- 15- 15- 10- 10- 10- 10- 10- 10- 10- 10
AdaBoost ◆ R2 State Model Name % correct Alpha ★ AdaBoost ◆ R2 Skin_CL_FLDRFS_1 78.5834 0.564583 ★ Skin_CL_FLDRFS_14 75.3794 0.130004 ★ Skin_CL_FLDRFS_8 74.1989 0.109510 ★ Skin_CL_FLDRFS_4 74.1989 0.476447 ★ Skin_CL_FLDRFS_45 74.0303 0.103133	-15

Skin_CL_FLDRFS_19 Skin_CL_FLDRFS_37 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_23 Skin_CL_FLDRFS_23 Skin_CL_FLDRFS_23 Skin_CL_FLDRFS_19 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_32 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_13 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_13 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_7 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_7 Skin_CL_FLDRFS_7 Skin_CL_FLDRFS_6 Skin_CL_FLDRFS_7 Skin_CL	Starting 593 C400 m0099 Un 73.8617 0.025789 0.203026 73.0185 0.203026 -0.001503 Cancel OK	Must negative same charge of atoms in the substru- Must positive sigma charge of atoms in the substruc Most positive sigma charge of atoms in the substruc Fast count the substructure (C3N)	Skin_PCA New parameter set name: Skin_PCA Recalculate ÖK Cancel
For help, press Fi		For Help, press Fi	

When a large number of sample data were applied to, 1. binary classification method --> perfect classification is not feasible. 2. PCA method --> clear sample plot and simple loading plot are difficult to generate.

When used in combination of the KY-method and PCA, this was generate great result. KY-method perfect classification is achieved.
 PCA method clear sample plot and simple loading plot are generated.

What is the KY-methods

Features of the KY (K-step Yard sampling) methods

- Always achieve perfect (100%) classification under any conditions

 Highly overlapped class sample data set
 Quite large number of sample set (tens and several thousands of)

 Starting sample set was divided into
- - small and clean sample set
- small and hierarchical sample set
 3. Applicable not only the discriminant but multi-regression analysis

"Two model KY-method for Discriminant analysis"

Classification results of compounds which are included in the "High reliability" space of the "KY-methods"





Total; 593, Positive 419, Negative; 174

Classification result by the KY-method (100% correct) Step1; Positive 187 Negative 88 Grey zone 318 Step2; Positive 177 Negative 34 Grey zone 107 Step3; Positive 55 Negative 52 Grey zone

Perfect classification(100%)

Classification Results by various methods (63params) Negative Methods Positive Total N.N. 85.6% 86.0% 86.2% SVM 91.7% 75.9% 98.3% LDA 87.0% 95.2% 67.2% KNN (K=5) 77.7% 55.8% 86.9% AdaBoost 82.1% 87.4% 80.0%

Conclusions:

1. Even if it was a large number of samples, it has been found that achieving an excellent data analysis results by using a combination of the KY-method and PCA.

2. Perfect (100%) classification was achieved by the KY-method 3. As a result of applying PCA for resampled sample set by the KY-method, it was possible to obtain a clean clustered sample space and much more simple and clear loading plot.

All research works and screen displays were executed and generated by the ADMEWORKS : ModelBuilder program developed by FJQS (Fujitsu Kyushu Systems Ltd.)