QSAR analysis of small datasets

Alexander Golbraikh & Alexander Tropsha

17th Euro-QSAR symposium
Uppsala, Sweden
September 22, 2008

OUTLINE

- *p*-values in statistical testing of classification QSAR models
- *p*-values in statistical testing of category QSAR models
- *p*-values in statistical testing of continuous QSAR models

What is a good classification QSAR model?

- **High correct classification rate:**

$$CCR = \frac{1}{K}\sum_{k=1}^{K} CCR_k = \frac{1}{K}\sum_{k=1}^{K} \frac{N_k^{corr}}{N_k^{total}} \geq Threshold$$

- **High correct classification rate for each class:**

$$CCR_k \geq Threshold_k, k = 1,2,...,K$$

For example: *Threshold*$_i$=0.70 can be considered acceptable

- **Statistical significance of prediction by QSAR models (usually is missing)**

  H_0 hypothesis: prediction is as good as random.

  H_1 hypothesis: prediction is better than random.

  With the given Level of Significance, can we reject H_0?

  Default Level of Significance: 0.95.

p-values in statistical testing of classification QSAR models

- **The main goal of classification QSAR:**
  build models with the highest classification accuracy for each class.
- **TEST SET:***
  Class 1: 8 compounds.
  Class 2: 4 compounds.
  Prediction by model:
  Class 1: 100% accuracy.
  Class 2: 75% accuracy.
- **H_0: Model predicts class 2 not better than random assignment of compounds to each class with equal probabilities.**
- **QUESTION: Can H_0 be rejected with the significance level of 95%?**

SOLUTION:

***p*-value for Class 2:** $p=0.5^4+4*0.5^4=5/16=0.31>>0.05$.

ANSWER: No.

*

p-values in statistical testing of classification QSAR models

APPROACH:

***p*-values for all classes should be below threshold (default: 0.05).**

Misclassification of one compound adds one to the total error.

***p*-value is calculated for the total error.**

Example 1:

Two classes.

Class 1: 10 compounds.

Total error: 3.

H_0: Equal probabilities in random assignment of a compound to each class.

QUESTION: Can H_0 be rejected?

SOLUTION:

The total number of ways 10 compounds can be distributed between two classes: 2^{10}=1024.

The number of ways 0 to 3 out of 10 compounds are misclassified:

0 compounds: 1

1 compound: 10

2 compounds: 45

3 compounds: 120

Total: 176.

***p*=176/1024=0.172>0.05.**

ANSWER: No.

p-values in statistical testing of classification QSAR models

Example 2:

Three classes.

Class 1: 10 compounds.

Total error: 3.

The number of ways N objects can be distributed between k classes with n_i objects in class i $(i=1,\ldots,k)$.

$$C = \frac{N!}{n_1!n_2!\ldots n_k!}, \sum_{i=1}^{k} n_k = N$$

H_0: Equal probabilities in random assignment of a compound to each class.

QUESTION: Can H_0 be rejected?

SOLUTION:

The total number of ways 10 compounds can be distributed between three classes: 3^{10}=59,049.

The number of ways 0 to 3 out of 10 compounds are misclassified:

- **All compounds are classified correctly: 1**
- **One compound is in class 2: 10.**
- **One compound is in class 3: 10.**
- **Two compounds are in class 2: 45.**
- **Two compounds are in class 3: 45.**
- **One compound is in class 2 and one compound is in class 3: 90.**
- **Three compounds are in class 2: 120.**
- **Three compounds are in class 3: 120.**
- **Two compounds are in class 2 and one compound is in class 3: 360.**
- **Two compounds are in class 3 and one compound is in class 1: 360.**

Total: 1161.

***p*=1161/59,049=0.0196<0.05.**

ANSWER: Yes.

p-values in statistical testing of classification QSAR models

Max error for which *p*-value<0.05 in case of 2 classes

Test set	Max error	Test set	Max error
1-4	-	37-39	13
5-7	0	40-41	14
8-10	1	42-43	15
11-12	2	44-46	16
13-15	3	47-48	17
16-17	4	49-50	18
18-20	5	60	23
21-22	6	70	27
23-25	7	80	32
26-27	8	90	36
28-29	9	100	41
30-32	10	200	87
33-34	11	500	231
35-36	12	**1000**	**473**

Why y-randomization test fails for small prediction sets?

If N→∞, Max.Error/N→0.5

AmpC beta-Lactamase HTS Assays*

- **278 models** used in consensus prediction of external evaluation set

CCR (Inhibitors;+)>70% for training and test sets

Test sets: 5-6 compounds
External Evaluation Set: 5 compounds
Models with FN=0: 204 (73.3%)
Models with TP=5 and FN=0: 49 (17.6%)
Test sets: statistically significant models: 12 (0 errors)
External Evaluation Set: 5 compounds
Models with FN=0: 12 (100%)
Models with TP=5 and FN=0: 4 (33.3%)

CCR (Nonbinders;-)>70% for training and test sets

Test sets: 5-8 compounds
External Evaluation Set: 5 compounds
Models with FP=0: 227 (81.7%)
Models with TN=5 and FP=0: 189 (68.0%)
Test sets: statistically significant models: 265
External Evaluation Set: 5 compounds
Models with FP=0: 219 (82.6%)
Models with TN=5 and FP=0: 181 (68.3%)

	FN=0	FN>0	Total
Sign.	12	0	12
Non-sign.	192	74	266
Total	204	74	278

***p*=2.23E-2**

	FP=0	FP>0	Total
Sign.	219	46	265
Non-sign.	8	5	13
Total	227	51	278

***p*=6.81E-2**

*Hsieh, J.H.; Wang, X.S.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput. Aided Mol. Des.* **2008,** 22, 593-609.

p-values in statistical testing of classification QSAR models

Max error for which *p*-value<0.05 in case of 3 classes

Test set	Max error	Test set	Max error
1-2	-	27-28	13
3-4	0	29	14
5-6	1	30-31	15
7-8	2	35	18
9-10	3	40	21
11-12	4	50	27
13-14	5	60	33
15	6	70	39
16-17	7	80	45
18-19	8	90	52
20-21	9	100	58
22	10	200	121
23-24	11	300	185
25-26	12	**500**	**315**

If N→∞, Max.Error/N→2/3

p-values

- **Theorem.** For any p-value $(0<p<1)$ and any $\varepsilon>0$, there exists number of objects N_p such that for any $N>N_p$ the maximum total error E_p of prediction corresponding to the p-value satisfies the condition $(E_p\text{-}E_{exp})/N<\varepsilon$, where E_{exp} is the expected total error.

- **Proof.** Let x be a random variable which takes values of 1 and 0 with probabilities q and $1\text{-}q$, respectively. Value of 1 means correct prediction and value of 0 means error. For each number N of experiments, $\boldsymbol{x}$ is binomially distributed with probability of success q**.** The probability of n successes out of N predictions is:

$$\Pr(n)=\binom{N}{n}q^n(1-q)^{N-n}$$

and the p-value is defined as follows:

$$p=\sum_{i=n}^{N}\Pr(n)=\sum_{i=n}^{N}\binom{N}{n}q^n(1-q)^{N-n}.$$

The expected value of successes is $N_{exp}=qN$, and the expected error is $E_{exp}=N(1\text{-}q)$. Suppose the number of predictions N increases. We assert that irrespective of p $(E_p\text{-}E_{exp})/N\rightarrow 0$ when $N\rightarrow\infty$.

In case of large N and n, the binomial distribution can be approximated by the normal distribution.

The density distribution function for the normal distribution is

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

$$\mu=qN \text{ and } \sigma=\sqrt{Nq(1\text{-}q)}$$

p-values

are its mean and standard deviation.

The cumulative distribution function is

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(x'-\mu)^2}{2\sigma^2}}dx'=\frac{1}{\sqrt{2\pi}\sigma}\left(1-\int_{x}^{\infty} e^{-\frac{(x'-\mu)^2}{2\sigma^2}}dx'\right)$$

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{x}^{\infty} e^{-\frac{(x'-\mu)^2}{2\sigma^2}}dx'=p$$

Here x is equal to the lowest number of successes corresponding to given p. Making a standard substitution $t=(x-\mu)/\sigma$, we obtain

$$\frac{1}{\sqrt{2\pi}}\int_{(x-\mu)/\sigma}^{\infty} e^{-\frac{t^2}{2}}dt=p$$

Since p does not change, $C=(x-\mu)/\sigma$ also does not change, i.e. it takes the same value for different N. Replacing μ and σ according to the above formulas, after simple transformations we obtain

$$x=qN-C\sqrt{Nq(1-q)},\ x/N=q-C\sqrt{q(1-q)/N}=N_{\exp}/N-C\sqrt{q(1-q)/N},$$

$$E_p/N=1-x/N=1-q+C\sqrt{q(1-q)/N}=E_{\exp}/N+C\sqrt{q(1-q)/N}$$

Thus, when $N\to\infty$, $(E_p-E_{exp})/N\to 0$.

p-values

2 classes, *p*-value<0.05, $q_1=q_2=0.5$, $E_{exp}=N/2$, $|E_p - E_{exp}|/N \to 0$

Test set,N	Max error, E_p	$\|E_p - E_{exp}\|/N$
37-39	13	0.149-0.167
40-41	14	0.150-0.159
42-43	15	0.143-0.151
44-46	16	0.136-0.152
47-48	17	0.138-0.146
49-50	18	0.133-0.140
60	23	0.117
70	27	0.114
80	32	0.100
90	36	0.100
100	41	0.090
200	87	0.065
500	231	0.038
1000	**473**	**0.027**

3 classes, *p*-value<0.05, $q_1=q_2=q_3= 1/3$, $E_{exp}=2N/3$, $|E_p - E_{exp}|/N \to 0$

Test set,N	Max error, E_p	$\|E_p - E_{exp}\|/N$
27-28	13	0.185-0.202
29	14	0.184
30-31	15	0.167-0.183
35	18	0.152
40	21	0.141
50	27	0.127
60	33	0.117
70	39	0.110
80	45	0.104
90	52	0.089
100	58	0.087
200	121	0.062
300	185	0.050
500	**315**	**0.037**

Confidence Intervals For Errors of Prediction

- **Classification problem with two classes.**

  Test set: N objects, k errors.

  Confidence intervals [α,β] for prediction of a large dataset.*

$$p(e_{true} \mid k, N) = \frac{p(k, N \mid e_{true})\, p(e_{true})}{p(k, N)}$$

$$p(e_{true} \mid k, N)\, de_{true} = \frac{p(k, N \mid e_{true})\, de_{true}}{p(k, N)} = e_{true}^{k} (1 - e_{true})^{N-k}\, de_{true}$$

$$\int_0^1 p(e_{true} \mid k, N)\, de_{true} = \int_0^1 e_{true}^{k} (1 - e_{true})^{N-k}\, de_{true} = B(k+1, N-k+1) = \frac{k!(N-k)!}{N!} = \frac{1}{C}$$

$$C \int_0^{\alpha} e_{true}^{k} (1 - e_{true})^{N-k}\, de_{true} = 0.025; \quad C \int_{\beta}^{1} e_{true}^{k} (1 - e_{true})^{N-k}\, de_{true} = 0.025$$

*Isaksson, A.; Wallman, M.; Göransson, H.; Gustafsson, M.G. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters, 2008, 29,1960-1965.

Confidence Intervals For Errors of Prediction

Statistically significant | **Not statistically significant**

Cmpds	Errors	α	β	Cmpds	Errors	α	β
5	0	0.004	0.460	3	0	0.006	0.603
6	0	0.003	0.410	4	0	0.005	0.522
7	0	0.003	0.370	3	1 (0-1)	0.067(0.011)	0.806(0.758)
8	0	0.002	0.337	4	1 (0-1)	0.052(0.009)	0.716(0.665)
8	1 (0-1)	0.027(0.004)	0.482(0.438)	5	1 (0-1)	0.042(0.007)	0.641(0.591)
9	0	0.002	0.309	6	1 (0-1)	0.036(0.006)	0.579(0.530)
9	1 (0-1)	0.024(0.004)	0.445(0.403)	7	1 (0-1)	0.031(0.005)	0.526(0.480)
10	0	0.002	0.285	8	2 (0-2)	0.074(0.006)	0.600(0.532)
10	1 (0-1)	0.022(0.003)	0.413(0.373)	9	2 (0-2)	0.066(0.005)	0.556(0.491)
11	0	0.002	0.265	10	2 (0-2)	0.059(0.004)	0.518(0.455)
11	1 (0-1)	0.020(0.003)	0.385(0.347)	10	3 (0-3)	0.108(0.006)	0.609(0.532)
11	2 (0-2)	0.054(0.004)	0.484(0.424)	11	3 (0-3)	0.098(0.005)	0.572(0.495)

What is a good prediction?

- **Two conditions should be satisfied:**

  1. low p-value (e.g. $p<0.05$)

  2. high accuracy (e.g. CCR≥0.70)

CONDITION 1 is critical for prediction of small number of compounds in the class.

CONDITION 2 is critical for prediction of larger number of compounds in the class.

Example:

Two classes.

H_0: Prediction is not better than random assignment of compounds to one of the two classes with probabilities $q_1=q_2=0.5$.

Starting from how many compounds in each class, CCR becomes lower than 0.70 while *p* is still lower than 0.05?

ANSWER: 23.

It means that starting from 23 compounds in the class, condition 2 is sufficient to consider the prediction acceptable.

If the class contains less than 23 compounds, condition 1 is sufficient to consider the prediction acceptable.

p-values in statistical testing of category QSAR models

Predicted classes	Observed classes				
	1	2	…	K	Total
1	N_{11}	N_{12}	…	N_{1K}	N_{1+}
2	N_{21}	N_{22}	…	N_{2K}	N_{2+}
…	…	…	…	…	…
K	N_{K1}	N_{K2}	…	N_{KK}	N_{K+}
Total	N_{+1}	N_{+2}	…	N_{+K}	N_{++}

CALCULATION OF ERRORS:

For Class 1: $(N_{21}-N_{11})+2(N_{31}-N_{11})+\ldots+(K-1)(N_{K1}-N_{11})$

For Class 2: $(N_{22}-N_{12})+(N_{32}-N_{22)=})+2(N_{42}-N_{22})+\ldots+(K-2)(N_{K2}-N_{22})$

etc.

p-values in statistical testing of category QSAR models

Example 1:

Three classes.

Class 1: 10 compounds.

Total error: 3.

H_0: Equal probabilities in random assignment of a compound to each class.

QUESTION: Can H_0 be rejected?

SOLUTION:

The total number of ways 10 compounds can be distributed between three classes: 3^{10}=59,049.

The number of ways 0 to 3 out of 10 compounds are misclassified:

- **All compounds are classified correctly: 1, error: 0.**
- **One compound is in class 2: 10, error: 1.**
- **One compound is in class 3: 10, error: 2.**
- **Two compounds are in class 2: 45, error: 2.**
- **One compound is in class 2 and one in class 3: 90, error: 3.**
- **Three compounds are in class 2: 120, error: 3.**

Total: 276.

P=276/59,049=4.67E-3<0.05.

ANSWER: Yes.

p-values

- **Theorem.** For any p-value $(0<p<1)$ and any $\varepsilon>0$, there exists N_p such that for any $N>N_p$ the maximum total error E_p of prediction corresponding to the p-value satisfies the condition $(E_p-E_{exp})/N<\varepsilon$, where E_{exp} is the expected total error.

- **Proof.** We consider a set of N compounds of class k. In case of K classes, let the probabilities of assigning a compound to each class be $q_1,q_2,\ldots,q_K$. Separately for each class, we consider assigning a compound to all other classes as success. We have already proven, that lowest numbers x_i of successes ($i=1,\ldots,K$) corresponding to given p

$$x_i / N = Q_i - C\sqrt{Q_i(1-Q_i)/N} = N^i_{\exp} / N - C\sqrt{Q_i(1-Q_i)/N}, Q_i = \sum_{j=1, j\neq i}^{K} q_j$$

$$E^i_p / N = e_{ki}(1-x_i)/N = e_{ki}(1-Q_i + C\sqrt{Q_i(1-Q_i)/N})$$

$$= e_{ki}(q_i + C\sqrt{Q_i(1-Q_i)/N}) = E^i_{exp} / N + e_{ki}C\sqrt{Q_i(1-Q_i)/N}$$

where $N^i_{\exp}$ is the expected number of compounds assigned to all classes except for class i and e_{kj} errors of assigning a compound of class k to class j ($e_{kk}=0$). If $N\to\infty$, the total error

$$E_p \to \sum_{j=1}^{K} E^i_{exp} = E_{exp}$$

and

$$(E_p - E_{exp})/N \to 0.$$

What is a good prediction?

- **Two conditions should be satisfied:**

  1. low p-value (e.g. $p<0.05$)

  2. high accuracy (e.g. CCR≥0.70)

CONDITION 1 is critical for prediction of small number of compounds in the class.

CONDITION 2 is critical for prediction of larger number of compounds in the class.

Example:

Three classes.

Category model.

Prediction is not better than random assignment of compounds to one of the three classes with $q_1=q_2=q_3=1/3$.

Starting from how many compounds in class 1, CCR becomes less than 0.70 while *p*-value is still lower than 0.05?

ANSWER: 8.

p-values in statistical testing of continuous QSAR models

- **Approach is based on the error of prediction.**

  Goal: calculate *p*-value for the given error of prediction.

Dataset: N compounds with observed activities x_i $(i=1,\ldots,N)$.

A QSAR model: Predicted activites y_i $(i=1,\ldots,N)$.

Assumptions: All y_i are within the interval $[a,b]$, where $a \le \min x_i$ and $b \ge \max x_i$.

For each compound *i*, an integrable probability distribution function $f_i(x)$ is defined on the interval $[a,b]$, which reaches maximum at $x=x_i$. Beyond the interval $[a,b]$, $f_i(x)$ is equal to zero.

$$\int_a^b f_i(x). = \int_{-\infty}^{\infty} f_i(x). = 1$$

Error of prediction: For each compound i, error of prediction is defined as $e_i=|y_i-x_i|$. The total error of prediction for the entire dataset is defined as sum of all e_i over all compounds:

$$E = \sum_{i=1}^{N} e_i = \sum_{i=1}^{N} |y_i - x_i|.$$

Maximum error of prediction is calculated as follows:

$$E_{\max} = \sum_{i=1}^{N} \max\{x_i - a, b - x_i\}.$$

p-values in statistical testing of continuous QSAR models

In following, we replace functions $f_i(x)$ by $f_i(x\text{-}x_i)=g_i(x)$ defined on the intervals $[a_i,b_i]$, where $a_i=a\text{-}x_i$ and $b_i=b\text{-}x_i$.

Procedure.

Let $r=E_{max}/K$, where K is some relatively large integer ~20-30 and define $q_1=r$, $q_2=2r,\ldots$, $q_K=Kr=E_{max.}$. The integral

$$p(q)=\int_0^{q} g_N(x_N)dx_N \int_0^{q-x_N} g_{N-1}(x_{N-1})dx_{N-1} \int_0^{q-x_N-x_{N-1}-\ldots-x_2} g_1(x_1)dx_1$$

gives the *p*-value for error q. Thus, the goal is to calculate this integral consecutively for q_1, q_2,... until its value exceeds 0.05 (or other defined threshold). If necessary, the value of the maximum allowed error could be determined more precisely by additional calculation of this integral with other *q*-values.

Example

If probabilities are uniformly distributed within the interval $[a,b]$, $p(q)$ can be calculated as follows.

$$p(q)=\frac{1}{(b-a)^N}\int_0^{q} g(x)dx_N \int_0^{q-x_N} g(x)dx_{N-1} \int_0^{q-x_N-x_{N-1}-\ldots-x_2} g(x)dx_1=\frac{1}{(b-a)^N}I(q),$$

where $$g(x)=\begin{cases}1, \text{if } x\in[a_i,b_i]\\ 0, \text{if } x\notin[a_i,b_i]\end{cases}.$$

p-values in statistical testing of continuous QSAR models

- **Example:**

  1 to 14 compounds randomly selected from the interval [1,5]. Minimum and maximum predicted activities are supposed to be within the interval [0,6]. 100 times.

  - **Maximum acceptable error of prediction based on *p*-value=0.05.**

Compounds	Total maximum error	Standard deviation	Mean max error per compound
1	0.15	-	0.15
2	0.90	-	0.45
3	2.00	0.016	0.67
4	3.19	0.064	0.80
5	4.48	0.11	0.90
6	5.78	0.14	0.96
7	7.18	0.18	1.03
8	8.53	0.24	1.07
9	9.92	0.25	1.10
10	11.29	0.31	1.13
11	12.69	0.31	1.15
12	14.23	0.36	1.19
13	15.71	0.41	1.21
14	17.13	0.41	1.22

p-values in statistical testing of continuous QSAR models

- **Example 1:**

  QSAR models were built for BBB permeability*.

  Test set (after excluding outliers): 10 compounds with activity values within the interval [-1.30, 1.44].

  Training set: 144 compounds with activity values within the interval [-2.15, 1.64].

  Consensus prediction of 10 compounds: total error (sum of errors for all 10 compounds): **2.95,** α=1.7E-5.

  Test set: Maximum error for p-value<0.05 and interval [-2.15, 1.64] : **7.30.**

  Prediction is statistically significant.

  (R^2=0.79 R_0^2=0.76 k=1.02 $R'_0{}^2$= 0.72 k'=0.73 F=29.8 α=6.0E-4)

- **Example 2:**

  MLR QSAR model was built for xanthone derivatives as α-glucosidase inhibitors**.

  Test set: 9 compounds with activity values within the interval [-2.37, -0.77].

  Training set: 34 compounds with activity values within the interval [-2.25, -0.76].

  Prediction of 9 compounds: total error: **2.40,** α=1.0E-3.

  Test set: Maximum error for p-value<0.05 and interval [-2.25, -0.76] : **3.80**; **1.10**

  Prediction is statistically significant.

  (R^2=0.82 R_0^2=0.76 k=1.07 $R'_0{}^2$= 0.48 k'=0.91 F=32.5 α=7.3E-4)

*Zhang, L.; et al. *Pharm Res.* **2008,** 25,1902-14.

Liu, Y.; et al. *Bioorg Med Chem*. **2008, 16, 7185-92.

Conclusions

- New statistical tests have been proposed to evaluate statistical significance of QSAR models.
- There exists a threshold value for a number of compounds of the prediction set, determining which statistical test should be used in evaluating QSAR models. If the number of compounds is below the threshold, *p*-value criterion should be used, otherwise classification accuracy criterion should be used. If the number of compounds is close to the threshold value, either one or another criterion should be used depending on the value of *p*-value.
- A new statistical criterion was developed to estimate predictive power of continuous QSAR models. This criterion can be used for small number of compounds in the prediction set.
- New software has been developed for most of these tests.

Acknowledgements

- **Prof. Clark Jeffries**
- **Dr. Georgiy Abramochkin**
- **Dr. Denis Fourshes**
- **Jui-Hua Hsieh**
- **Kun Wang**
- **Liying Zhang**
- **Chris Grulke**
- **Hao Tang**

The project was supported by the following grants:

P20-HG003898. **National Institutes of Health**
Carolina Exploratory Center for Cheminformatics Research (NIH Roadmap).

R01-GM66940-06A1. **National Institutes of Health**
Predictive QSAR Modeling.