Applied Medical Informatics
Original Research Vol. 28, No. 2 /2011, pp: 9-14

Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique

Sorana D. BOLBOACĂ[1,*]

[1] "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, 400023 Cluj-Napoca, Romania.
E-mail: sbolboaca@umfcluj.ro

* Author to whom correspondence should be addressed; Tel.: +4-0264-431697; Fax: +4-0364-818418.

Received: 1 May 2011 /Accepted: 27 May 2011 / Published online: 15 June 2011

Abstract
Aim: The properness of random assignment of compounds in training and validation sets was assessed using the generalized cluster technique. *Material and Method*: A quantitative Structure-Activity Relationship model using Molecular Descriptors Family on Vertices was evaluated in terms of assignment of carboquinone derivatives in training and test sets during the leave-many-out analysis. Assignment of compounds was investigated using five variables: observed anticancer activity and four structure descriptors. Generalized cluster analysis with K-means algorithm was applied in order to investigate if the assignment of compounds was or not proper. The Euclidian distance and maximization of the initial distance using a cross-validation with a v-fold of 10 was applied. *Results*: All five variables included in analysis proved to have statistically significant contribution in identification of clusters. Three clusters were identified, each of them containing both carboquinone derivatives belonging to training as well as to test sets. The observed activity of carboquinone derivatives proved to be normal distributed on every. The presence of training and test sets in all clusters identified using generalized cluster analysis with K-means algorithm and the distribution of observed activity within clusters sustain a proper assignment of compounds in training and test set. *Conclusion*: Generalized cluster analysis using the K-means algorithm proved to be a valid method in assessment of random assignment of carboquinone derivatives in training and test sets.

Keywords: quantitative Structure-Activity Relationship (qSAR); Molecular Descriptors Family on Vertices (MDFV); Anticancer drug; Generalized Cluster Analysis.

Introduction

Cluster analysis techniques [1] comprise different algorithms able to group similar objects. The most known algorithms are: k-means algorithm (Forgy's algorithm [2]) and its variants (trimmed k-means algorithm [3], fast-MCD algorithm [4], bisecting k-means algorithm [5], Principal Direction Divisive Partitioning (PDDP) algorithm [6]), hierarchical algorithm (e.g. agglomerative [5], divisive [7], BIRCH [8]) (displayed graphically using cluster-subcluster relationships with a dendrograms), Density-Based Clustering algorithm [9], etc.

Clusterization techniques are widely used in many research fields. In quantitative structure-activity / structure-property relationships (qSARs / qSPRs), mathematical approaches able to link the structure of compounds with activity/property [10], the cluster methods are used as methods

for selecting training and validation sets. Hierarchical cluster based methods (single linkage, complete linkage, group average, Wards method, centroid method and median method) or non-hierarchical methods (K-means [2], Jarvis-Patrick clustering [11], DBSCAN [12], OPTICS [13], DENCLUE [14]) are used beside random selection, Kohonen's self-organizing map and informative design [15] in inclusion of compounds in training and test sets [16].

A sample of 37 carboquinone derivatives was previously investigated using self-organizing map to qSAR analysis [17]. Their model proved to have a predictive ability with an average of 4.2% error and a cross-validation correlation coefficient of 0.87. This set of carbiquinone derivatives was also investigated using Molecular Descriptors Family on Vertices [18] and this model proved to be performing compared to the other investigated models. The qSAR MDFV models of carboquinone derivatives was used in the present researcher in order to assess if the applied random assignment of compounds in training and test sets was or not a proper method to split the compounds in the training and test sets.

Material and Method

qSAR MDFV Model

The qSAR model with four MDFV variables previously obtained [18] was investigated in this research. The model is presented in Eq(1).

$$\hat{Y} = 24.26(\pm 4.32) - TEuIFFDL*2.40(\pm 0.47) - GLCIicdI*16.78(\pm 4.38) - TAkaFcDL*0.65(\pm 0.11) - GlbIAcDR*0.02(\pm 0.01) \quad (1)$$

where $\hat{Y}$ = anticancer activity estimated by model from Eq(1), TEuIFFDL, GLCIicdI, TAkaFcDL and GlbIAcDR = values of MDFV members.

The statistical characteristics of the model from Eq(1) are presented in Eq(2)

$r^2 = 0.9548$; $s_{est} = 0.14$; n = 37; F-value = 169 (p-value = $5.01 \cdot 10^{-21}$);
r (p-value) = 0.9771 ($4.07 \cdot 10^{-25}$); ϱ (p-value) = 0.9461 ($1.3 \cdot 10^{-18}$);
r_{sQ} (p-value) = 0.9615 ($3.26 \cdot 10^{-21}$); τ_a (p-value) = 0.8273 ($5.74 \cdot 10^{-13}$) = τ_b; (2)
τ_c (p-value) = 0.8050 ($2.35 \cdot 10^{-12}$); Γ (p-value) = 0.8361 ($1.13 \cdot 10^{-9}$);
$r^2_{loo} = 0.9351$; $s_{pred} = 0.17$; F_{pre} (p-values$_{pred}$) = 115 ($5.42 \cdot 10^{-20}$)

where est = estimated; pred = predicted; n = sample size; r = Pearson correlation coefficient [19]; r^2 = determination coefficient; s_{est} = standard error of estimate; F-value = Fisher statistic of the MLR model; ϱ = Spearman rank correlation coefficient [20]; r_{sQ} = semi-quantitative correlation coefficient [21]; $\tau_{a,b,c}$ = Kendall tau a, b and c correlation coefficients [22]; Γ = Gamma correlation coefficients [23]; r_{loo}^2 = determination coefficient obtained in leave-one-out analysis; s_{pred} = standard error of predicted; F_{loo} = F-value obtained in leave-one-out analysis.

The qSAR with identified MDFV variables (Eq(1)) was also tested by applying the leave-many out analysis. The set was randomly spit in training and test set with ~ 1/3 of compound in the test set and in respects of normality of experimental data. Based on the descriptors used by Eq(1), a model was identified using compounds assigned to training set and was validated on compounds assigned to test set (cqd03, cqd04, cqd06, cqd09, cqd11, cqd18, cqd21, cqd22, cqd23, cqd26, cqd27, and cqd37). The model obtained in training set and its statistical characteristics of model obtained in training set is presented in Eq(3). Statistical characteristic of model in test set is presented in Eq(4).

$$\hat{Y} = 21.582 + -2.4660*TEuIFFDL + -14.253*GLCIicdI + -6.2922e\text{-}1*TAkaFcDL + -0.0217*GLbIAcDR \quad (3)$$

Training set: $r^2 = 9480$, F-value = 82, p-value = $2.67 \cdot 10^{-11}$

Test set: $r^2 = 0.9675$, F-value = 38, p-value = $1.24 \cdot 10^{-5}$ (4)

Assessment of Random Selection

Generalized cluster analysis with K-means algorithm was applied in order to investigate the

qSAR MDFV model of carboquinone derivatives. The Euclidian distances (as distance measure, measure how fat two values are from each other, similar values are identified by a low distance) and maximization of the initial distance (in regards of cluster center) were the criterion applied. A cross-validation with a v-fold of 10 was applied. The assessment was carried out using STATISTICA 8.

Results

The generalized cluster analyses, applied using five quantitative variables (anticancer activity and the values of MDFV variables from Eq(1)), identified 3 clusters. The training error proved to be of 0.2812 (n=37). Statistical parameters associated to variables included in analysis according to the assigned cluster are presented in Table 1.

Table 1. Centroids for K-means clustering

Cluster	log(1/C)	TEuIFFDL	GLCIicdI	TAkaFcDL	GLbIAcDR	No (no.tr:no.ts)	% (%tr:%ts)
1	4.9670	0.1338	0.9867	1.9231	45.3238	10 (7:3)	27 (70:30)
2	6.1150	34.0550	-0.0197	0.9913	1.3480	2 (1:1)	5 (50:50)
3	6.0416	-0.0399	0.9800	1.2124	44.3843	25 (17:8)	68 (68:32)

TEuIFFDL, GLCIicdI, TAkaFcDL, GLbIAcDR = MDFV descriptors used by Eq(1); tr = training; ts = test

The assignment of compounds in clusters, according to training and test set, is presented in Table 2.

Table 2. Generalize cluster analysis: assignment of compounds in clusters

Cluster	Set	Compounds
1	Training	cqd01, cqd02, cqd05, cqd07, cqd08, cqd13, cqd19
	Test	cqd03, cqd04, cqd06
2	Training	cqd35
	Test	cqd18
3	Training	cqd10, cqd12, cqd14, cqd15, cqd16, cqd17, cqd20, cqd24, cqd25, cqd28, cqd29, cqd30, cqd31, cqd32, cqd33, cqd34, cqd36
	Test	cqd09, cqd11, cqd21, cqd22, cqd23, cqd26, cqd27, cqd37

The distance to centroid varied from 0.0890 to 0.5403 for first cluster and from 0.0908 to 0.5289 for the third cluster. Normalized mean of anticancer activity and MDFV values are presented in Figure 1 while statistical characteristics of variables included in each cluster are presented in Table 3.

Table 3. Generalized cluster analysis k-means: results

Parameter	Cluster 1 (n=10)		Cluster 2 (n=2)		Cluster 2 (n=25)	
	Mean	StDev	Mean	StDev	Mean	StDev
log(1/C)	4.9670	0.4667	6.1150	0.6010	6.0416	0.3888
TEuIFFDL	0.1338	0.1120	34.0550	0.0636	-0.0399	0.0898
GLCIicdI	0.9867	0.0130	-0.0197	0.1567	0.9800	0.0108
TAkaFcDL	1.9231	0.3952	0.9913	0.0123	1.2124	0.3890
GLbIAcDR	45.3238	11.3824	1.3480	0.2547	44.3843	9.4494

StDev = standard deviation

The contribution of each variable to the assignment of carboquinone derivatives in the clusters is presented in Table 4.

The distribution of anticancer activity according to the cluster of the carbocquinone derivatives is presented in Figure 2.

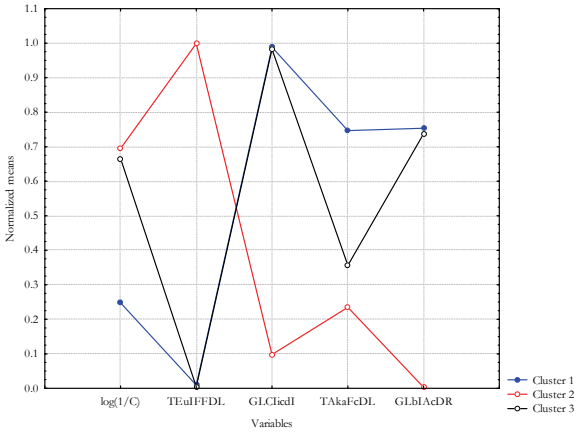

Figure 1. Graph of means for continuous variables

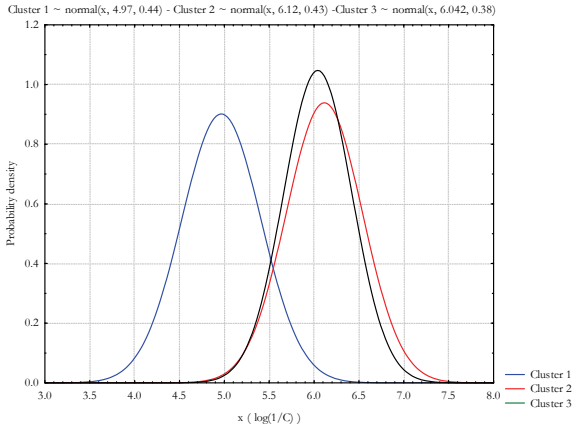

Figure 2. Distributions of log(1/C) variable

Table 4. Generalized cluster analysis k-means: results of variance analysis

Variable	Between SS[a]	Within SS[b]	F-value	p-value
log(1/C)	1228.27	5.950	3509.6	< 0.0001
TEuIFFDL	2317.85	0.310	126906.5	< 0.0001
GLCIicdI	32.82	0.029	19326.0	< 0.0001
TAkaFcDL	74.30	5.038	250.7	< 0.0001
GLbIAcDR	69752.99	3309.106	358.3	< 0.0001

SS = sum of squares; [a] df (degrees of freedom) = 2; [b] df = 34;
F-value = Fisher statistics; p-value = Fisher's significance

Discussion

Generalized cluster analysis with K-means algorithm was successfully applied to investigate the assignment of carboquinone derivatives in the training and test sets. Three clusters were identified using this technique when the Euclidian distance and maximization of the initial distance on a cross-validation of v-fold of 10 were the criteria used. The analysis of the centroid for K-means clustering (Table 1) revealed the following:

- Three variables proved to have distinct values in the identified clusters: log(1/C), TEuIFFDL and TAkaFcDL MDFV descriptors.
- The second cluster proved to have distinct values of all variables used in generalized cluster analysis.
- Similar centroids could be identified for first and third cluster in regards of GLCIicdI and GLbIAcDR MDFV descriptors.
- Compounds from both training and test sets could be identified among identified clusters.
- Almost 30% of compounds in clusters with more than 2 compounds proved to be from test set. The exception is represented by second cluster that comprise just two carboquinone derivatives, one from training set and the other from test set.

The analysis of Figure 1 and of statistical characteristics of variable for each cluster (Table 3) revealed the following:

- First cluster comprised carboquinone derivative with small log(1/C) and high value of MDFV structure descriptors.
- Second cluster comprised carboquinone derivative with high log(1/C) and of TEuIFFDL descriptors as well as small values of all other MDFV structure descriptors.
- Third cluster comprised carboquinone derivative with high log(1/C), similar values of TEuIFFDL and GLCIicdI to the first cluster, intermediary values of TAkaFcDL and high value of GLbIAcDR (value similar to the contribution of GLbIAcDR to the first clusters).

All variables (log(1/C) and four MDFV members) proved to have statistically significant contributions to assignment of compounds in the clusters (all p-values < 0.0001 – Table 4). This result showed that all variables contribute significantly statistics to clusterization. Moreover, the observed activity of carboquinone derivatives proved to be normal distributed on each cluster while the imposing criterion in random splitting of compounds in training and test set was the normality of observed activity in both sets.

The presence of training and test sets in all clusters identified using generalized cluster analysis with K-means algorithm and the distribution of observed activity within clusters sustain a proper assignment of compounds in training and test set.

Cluster analysis techniques are used in qSAR as well as in virtual screening for identification of active compounds. These techniques are used to identify similar structural feature among compounds [24], new active compounds [25], to select descriptors [26], to split the compounds in training and validation sets [27], etc. These techniques are used [28-30] despite theirs disadvantages: different methods provide very different results (since different criteria for merging clusters are used), the results could be affected when ordered variables are used, unstable results when cases are withdrawn, etc. In the current manuscript, the K-means algorithm was used to assess a random split of compounds in training and test sets and was successfully accomplished. The generalized cluster analysis with K-means algorithm proved its usefulness in assessment of random assignment of carboquinone derivatives compounds in training and validation sets. It is expected to observe the same results when the generalized cluster analysis with K-means algorithm is applied on other qSAR models. However, a study is currently carried out in order to validate the result obtained on the carboquinone derivative set.

Conclusion

The generalized cluster analysis with K-means algorithm proved reliable method for assessment of proper assignment of carboquinone derivatives in training and test sets in the leave-many-out analysis in a quantitative Structure-Activity Relationship experiment.

Conflict of Interest

The author declares that there is no conflict of interest.

Acknowledgements

Financial support is gratefully acknowledged to UEFISCSU Romania (ID0458/206/01.10.2007). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Tryon RC. Cluster Analysis. Ann Arbor, MI: Edwards Brothers, 1939.
2. Forgy EW. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 1965;21:768-769.
3. García-Escudero LA, Gordaliza A, Matrán C. Trimming tools in ex-ploratory data analysis. J Comput Graph Statist 2003;12:434-449.
4. Rousseeuw PJ, Van Driessen K. A fast algorithm for the Minimum Co-variance Determinant estimator. Technometrics 1999;41:212-223.
5. Jain AK, Murty MN, Flynn PJ. Data clustering: a Review. ACM Comput Surv 1999;31(3):264-323.
6. Boley DL. Principal Direction Divisive Partitioning. Data Min Knowl Disc 1998;2(4):325-344.
7. Sneath PHA, Sokal RR. Numerical Taxonomy. Freeman: London, UK, 1973.
8. Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large

databases. SIGMOD Rec 1996;25(2):103-114.
9. Jain AK, Dubes RC. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. 1988.
10. Hammett, L.P. Some Relations between Reaction Rates and Equilibrium Constants. Chem Rev 1935;17(1):125-136.
11. Jarvis RA, Patrick EA. Clustering using a similarity measure based on shared near neighbours. IEEE Transactions in Computers 1973;C-22:1025-1034.
12. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining 1996:226-231.
13. Ankrest M, Breunig M, Kriegel H, Sander J. OPTICS: Ordering points to identify the clustering structure. Proceedings of the ACM SIGMOD International Conference on Management of Data 1999:49-60.
14. Han JW, Kamber M. Data mining: concepts and techniques. San Francisco, Morgan Kaufmann Publishers. 2001.
15. Miller JL, Bradley EK, Teig SL. Luddite: An information-theoretic library design tool. J Chem Inf Comp Sci 2002;43(1):47-54.
16. Tsygankova IG. Variable Selection in QSAR Models for Drug Design. Curr Comput Aided Drug De 2008;4(2):132-142.
17. Kawakami J, Hoshi K, Ishiyama A, Miyagishima S, Sato K. Application of a self-Organizing Map to Quantitative Structure-Activity Relationship Analysis of Carboquinone and Benzodiazepine. Chem Pharm Bull 2004;52(6):751-755.
18. Bolboacă SD, Jäntschi L. Comparison of QSAR Performances on Carboquinone Derivatives. TheScientificWorldJOURNAL 2009;9(10):1148-1166.
19. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Phil Trans R Soc A 1896;187:253-318.
20. Spearman C. General intelligence" objectively determined and measured. Am J Psychol 1904;15:201-293.
21. Bolboacă SD, Jäntschi L. Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. Leonardo J Sci 2006;9:179-200.
22. Kendall MG. A New Measure of Rank Correlation. Biometrika 1938;30:81-89.
23. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29:1-27.
24. Rescigno A, Casañola-Martin GM, Sanjust E, Zucca P, Marrero-Ponce Y. Vanilloid Derivatives as Tyrosinase Inhibitors Driven by Virtual Screening-Based QSAR Models. Drug Test Anal 2011;3(3):176-181.
25. Hecht D. Applications of machine learning and computational intelligence to drug discovery and development. Drug Develop Res 2011;72(1):53-65.
26. Le-Thi-Thu H, Cardoso GC, Casañola-Martin GM, Marrero-Ponce Y, Puris A, Torrens F, Rescigno A, Abad C. QSAR models for tyrosinase inhibitory activity description applying modern statistical classification techniques: A comparative study. Chemometr Intell Lab 2010;104(2):249-259.
27. Yu Y, Su R, Wang L, Qi W, He Z. Comparative QSAR modeling of antitumor activity of ARC-111 analogues using stepwise MLR, PLS, and ANN techniques. Med Chem Res 2010;19(9):1233-1244.
28. Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design - A review. Curr Top Med Chem 2010;10(1):95-115.
29. Liu W, Johnson DE. Clustering and its application in multi-target prediction. Curr Opin Drug Discov Devel 2009;12(1):98-107.
30. Hecht D. Applications of machine learning and computational intelligence to drug discovery and development. Drug Develop Res 2011;72(1):53-65.