Repeated double cross validation for optimization and evaluation of empirical classifiers

Kurt Varmuza[1], Bettina Liebmann[1], Peter Filzmoser[2]

1 Institute of Chemical Engineering, Vienna University of Technology, Vienna, Austria, www.lcm.tuwien.ac.at
2 Institute of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria, www.statistik.tuwien.ac.at/public/filz

Classification of objects

Since the very beginning of chemometrics, classification of objects has been an important task. Common deficiencies of empirical classification models (classifiers) are

- inappropriate performance measures,
- poor strategies for (1) optimizing the complexity of classifiers , (2) estimating the performance of classifiers,
- mixing model optimisation and performance estimation.

We present the powerful and easily applicable strategy **repeated double cross validation** (rdCV) [1, 2], and apply rdCV to the classification methods (1) D-PLS, discriminant PLS, (2) KNN, *k*-nearest neighbour classification, and (3) SVM, support vector machine classification.

Multivariate classification methods

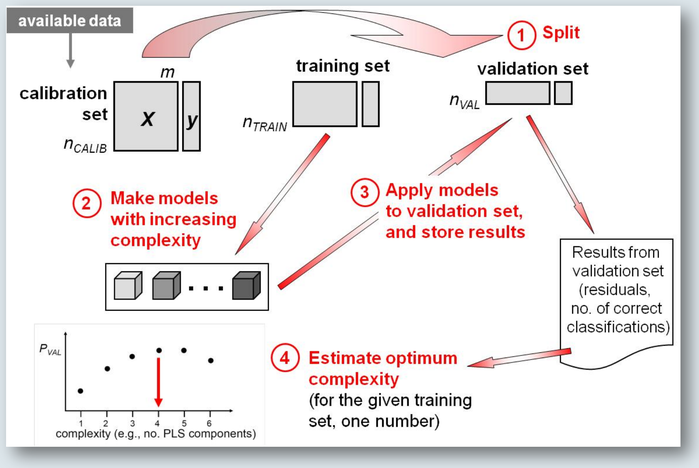
D-PLS, Discriminant PLS. Linear, binary classification of class 1 (y=-1), and class 2 (y = +1) . Class assignment: if $\hat{y} < 0$ assign to class 1, else to class 2.
Optimisation of model complexity: *a*, number of PLS components.

KNN, *k*-nearest neighbour classification. Nonlinear classification based on the (Euclidean) distance between objects in *x*-space. Find nearest neighbours (objects with known class membership) to query object.
Optimisation of model complexity: *k*, number of neighbours.

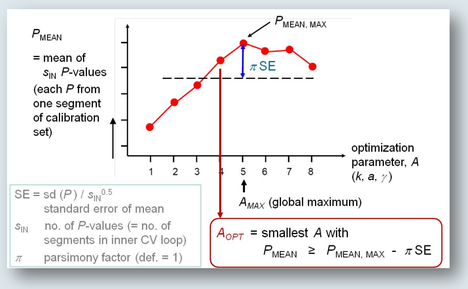
SVM , Support Vector Machine classification. Nonlinear classification.
Optimisation of SVM parameter γ.

Repeated double cross validation, rdCV

Optimum model complexity

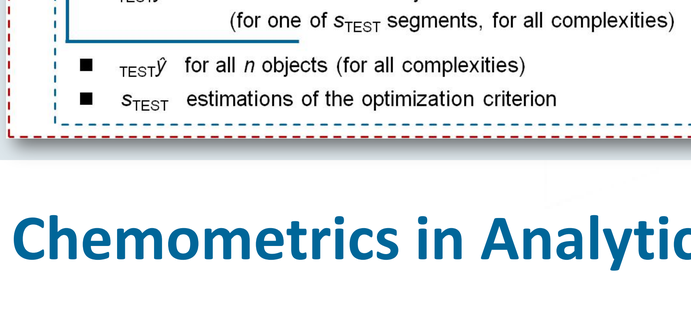


"Standard error method"



The standard error of the mean, SE, is used to find the optimum optimization parameter, A_{OPT} [3].

Performance for new cases



Depends on (random) split into calibration set and test set.
Thus, repetitive random splits are desirable!

"Predictive ability *P*"

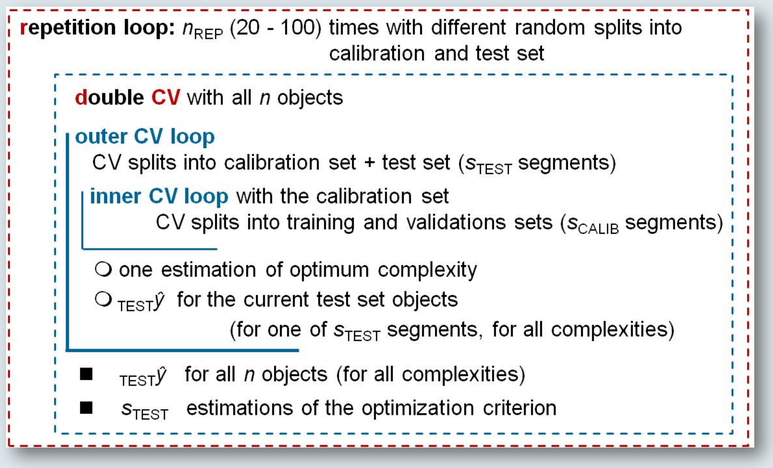
Class assignment table (binary classification)

true class \ assigned class	1	2	sum
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2
sum	$n_{\to 1}$	$n_{\to 2}$	n

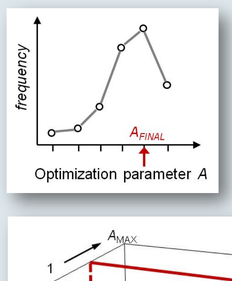
Predictive ability class 1: $P_1 = n_{11}/n_1$
class 2: $P_2 = n_{22}/n_2$
Average predictive ability: $P = (P_1 + P_2)/2$

Avoid:
Overall predictive ability = $(n_{11} + n_{22})/n$
It is very misleading if classes differ much in size.

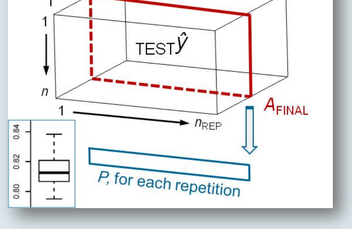
rdCV scheme



rdCV results



Optimum model complexity estimated from calibration data
(only using inner CV loops)
From n_{REP} x s_{TEST} optimisations of the optimum complexity *A*, the most frequent value is taken as the final optimum complexity A_{FINAL}.

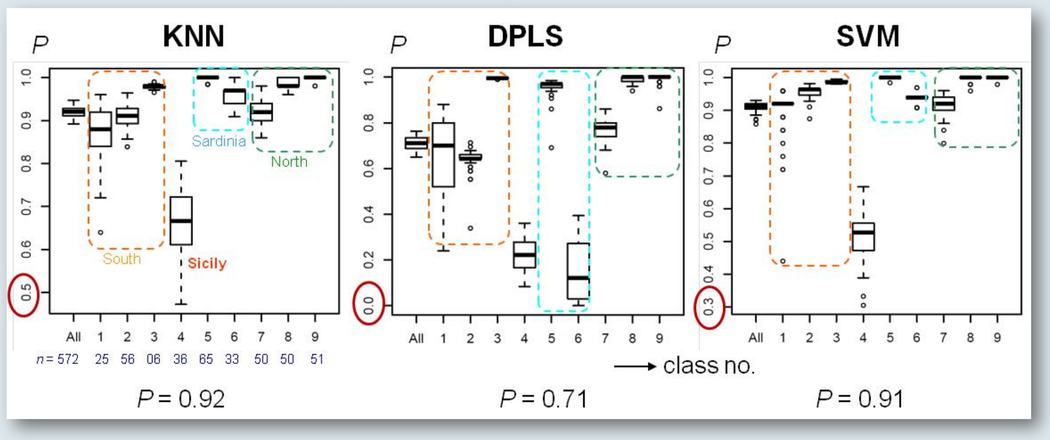


Performance for new cases estimated from test data
(only using outer CV loops)
From *n* x n_{REP} test set predictions at A_{FINAL}, the predictive ability *P* (or other performance criteria) is calculated for each repetition.
The variation of *P* is shown in a box plot.

Example 1: Origin of Italian olive oil

n = 572 oils, 9 classes (with 25 to 206 samples) from 9 areas in Italy, *m* = 8 fatty acid concentrations [4], R [5] package "classify", data(olive).

rdCV: n_{REP} = 50 repetitions, s_{OUT} = 2, s_{IN} = 6
Optimization results: KNN: k_{FINAL} = 1; DPLS: a_{FINAL} = 7; SVM: γ_{FINAL} = 0.07

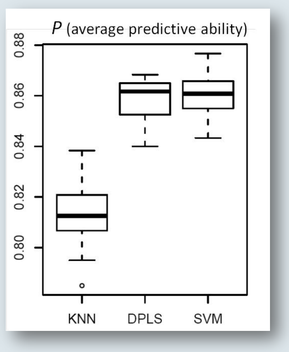


Both KNN and SVM show good prediction performance; the mean of their average predictive ability is around 0.9, except for the oils from class "Sicily". For this multiclass problem, DPLS performs worse.

Example 2: Spectra-structure relationship

Chemical substructures present (class 1) / not present (class 2), *n* = 600 (class 1: 300, class 2: 300), *m* = 658 spectral descriptors derived from mass spectra [6], R-package "chemometrics" [7], data(phenyl)

rdCV: n_{REP} = 20 repetitions, s_{OUT} = 2, s_{IN} = 6
Optimization results: KNN: k_{FINAL} = 3; DPLS: a_{FINAL} = 2; SVM: γ_{FINAL} = 0.0002
Computation time: KNN 550 s, DPLS 42 s, SVM 940 s.



For this 2-class problem, DPLS performs equally well as SVM classification. KNN classification shows lower average predictive ability and more variation of *P*.

Conclusions

- rdCV is a resampling method combining some systematics and randomness.
- rdCV is applicable to calibration and classification problems for data sets with approximately ≥ 25 objects.
- In rdCV, optimization of model complexity (model parameter) is separated from the estimation of model performance.
- rdCV provides estimations of the variability of model complexity and of model performance.
- rdCV is easily applicable, fast and free: R-package "*chemometrics*" [7], www.lcm.tuwien.ac.at/R

References

[1] P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometrics, 23, 160 (2009).
[2] K. Varmuza, P. Filzmoser: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009).
[3] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd ed., Springer, NY, USA (2009).
[4] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia: in Food Research and Data Analysis; ed. H. Martens, H. Russwurm Jr., Applied Science Publ. London, 189-214 (1983)
[5] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2011. www.r-project.org
[6] W. Werther, W. Demutz, F.R. Krueger, J. Kissel, E.R. Schmid, K. Varmuza, J. Chemom, 16, 99 (2002).
[7] P. Filzmoser, K. Varmuza: chemometrics: Multivariate Statistical Analysis in Chemometrics, R-package v. 1.3.8, 16-02-2012. http://cran.r-project.org/web/packages/chemometrics/index.html

Acknowledgements. Anton Friedl (Institute of Chemical Engineering, TU Vienna, Austria) is warmly thanked for providing financial support for the presenting author.