

PLS-trees in data mining and clustering

Lennart Eriksson¹, Johan Trygg² and Svante Wold³

¹ Umetrics AB, POB 7960, SE-907 19 Umeå, Sweden, lennart.eriksson@umetrics.com

² Institute of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, johan.trygg@chem.umu.se

³ Umetrics Inc., 17 Kiel Ave., Kinnelon, NJ 07405, USA, svante.wold@umetrics.com

Keywords: PLS-trees, dendrogram, data mining, clustering, outlier detection.

1 Introduction

To find relationships in large data sets, such data sets are almost always divided into groups containing fairly homogeneous (non-grouped) data. Hence, to make multivariate data analysis applicable in data mining, and other analyses of large and complex data sets, we need one or several clustering algorithms, preferably combined with an embedded multivariate regression step (i.e. PLS, OPLS® or O2PLS®). This multivariate clustering algorithm must work well for large data sets with potentially very many and collinear variables, missing data, noise, and other common complications.

The objective of this contribution is to present a top-down hierarchical clustering approach based on a set of connected PLS models. Called PLS-trees, this approach is analogous to classification and regression trees (CART), but uses the scores of PLS regression models as the basis for splitting the clusters, instead of the individual X-variables. When applied to a pair of matrices, X and Y, the approach results in the row-wise splitting of data into a tree structure (dendrogram) of PLS models, one split for each cluster (node in the dendrogram). Figure 1 shows an example dendrogram, extracted from SIMCA-P+® 12. The dendrogram with the associated PLS models is called a PLS-tree. When Y comprises a discrete matrix with 1/0 columns corresponding to a number of predefined classes, the result is a PLS classification tree.

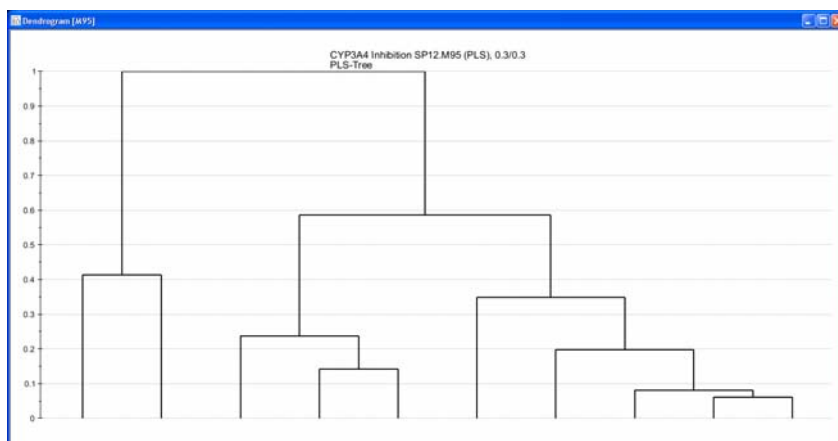


Figure 1. A typical PLS-tree. Each branch of the tree corresponds to a PLS model fitted to a sub-set of observations.

2 Theory

Briefly, the PLS-tree approach works as follows: The split of one cluster into two is made along the sorted first X-score (t_1) of a cluster PLS model. The position of the split along the score is selected according to the improvement of a combination of (a) the variance of the X-score (t_1), and (b) the variance of Y, and (c) a penalty function encouraging a balanced split with approximately equal numbers of observations in each resulting branch. Cross-validation is used to terminate the branches of the tree, and to determine the number of components of each cluster PLS model. To accomplish a PLS-tree the data matrices, X and Y, are first centered and scaled, as usual. A PLS analysis is made of X and Y. Thereafter, the first X-score (t_1) is used as the dividing coordinate together with the Y-data (X and Y are sorted along t_1). The point on t_1 is searched that divides X and Y in two parts, 1 and 2, such that the following is minimized:

$$B * F(N_1, N_2) + (1-B) * [A (V_{Y,1} + V_{Y,2}) + (1-A) (V_{t_1,1} + V_{t_1,2})]$$

In the expression above, V denotes variance, F is a function of the number of observations in the sub-groups 1 and 2, and A and B are two adjustable parameters. F is large when N_1 is very different from N_2 . The parameters A and B both run between 0 and 1. They regulate how the PLS models are split (i.e. how the observations from one upper-level PLS model are distributed among two lower-level models) according to the score t_1 , the Y -variable(s) or the group size. The first parameter, A , sets the balance between the score t_1 and the Y ; the closer to zero, the more weight is attributed to the score t_1 . The second parameter, B , takes into account the group size of the resulting clusters; the closer to zero, the less important it becomes to have equal group sizes in the dendrogram. In summary, this means that a division along t_1 is sought that minimizes the within group variation and hence maximizes the between group differences in t_1 and Y .

3 Material and methods

The first application data set is a QSAR data set and concerns the inhibition of CYP3A4 for a series of 930 compounds [1]. A training set of 551 compounds was defined by means of onion design. The corresponding prediction set therefore comprises 379 compounds. In order to account for the physico-chemical properties of the compounds, a total of 307 chemical descriptors were assembled. The biological effect is the $\log IC_{50}$ to the enzyme. This data set is fairly homogenous with uniform coverage properties of the PLS score space (Figure 2). The objective of analyzing this data set is to investigate whether the PLS-tree approach may discover subtle groupings in the training set. Such possible sub-groupings may correspond to locally preserved themes in the SAR, and may warrant more than one PLS model.

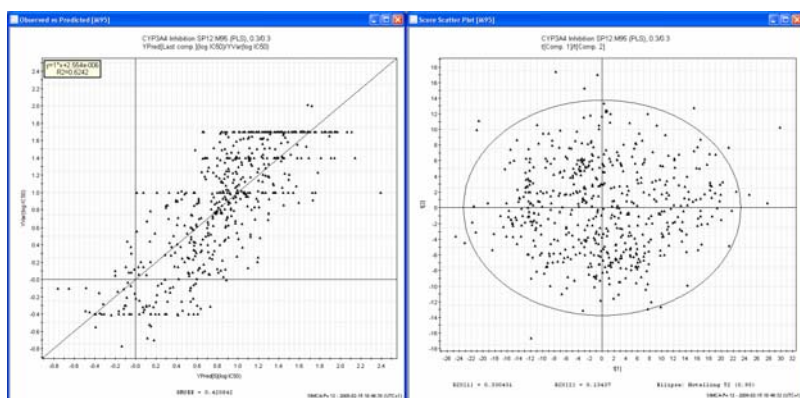


Figure 2. Plots of the reference PLS model based on the training set of 551 compounds. (left) Scatter plot of observed and predicted biological effects. (right) Scatter plot of the first two scores, t_1 and t_2 .

The second illustration is also a QSAR data set and deals with soil sorption of environmental pollutants [2]. In contrast to the first example, this data set exhibits pronounced clustering in the PLS score space (Figure 3). The $\log KOC$ data set contains 351 compounds distributed across 14 known chemical classes. There are 64 chemical descriptors. The environmental effect is the soil sorption coefficient, denoted $\log KOC$. The objective of analyzing the second data set is to investigate whether the PLS-tree approach may confirm existing clustering, whether some clusters can be united to larger groups, and if local QSAR modeling may lead forward to enhanced predictive ability.

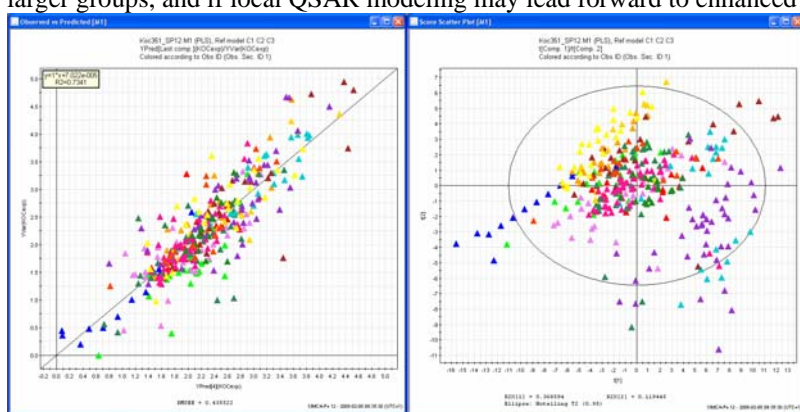


Figure 3. (left) Agreement between observed and predicted $\log KOC$ for the PLS reference model. (right) Score plot t_1/t_2 . In both plots the coding is done according to chemical class. The clustered nature of the data set is obvious.

4 Results and discussion

In the analysis of the first data set, it was decided to vary A and B using a 3^2 factorial design with the settings 0.1, 0.3 and 0.5 in A and B . Nine PLS-trees were initiated using the values of A and B . Figure 4 shows the dendrograms.

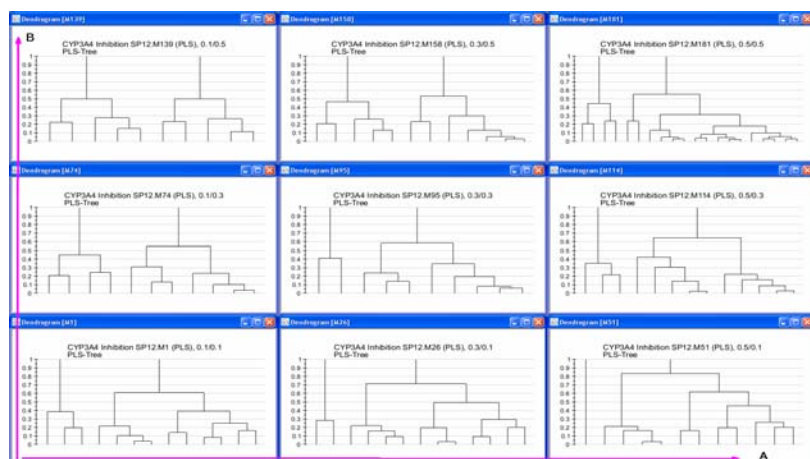


Figure 4. The nine PLS-trees obtained by fitting PLS models to the training set of 551 compounds and using different values of the parameters A and B, as dictated by a 3^2 factorial design.

As shown by Figure 4, the dendrograms have quite a varied appearance. Some features are pretty “predictable” and some are more “unexpected”. For instance, the higher the value of B the more equal size the resulting groups will tend to have. The top left dendrogram in Figure 4 (based on $A=0.1$ and $B=0.5$) clearly visualizes the impact of a high B, i.e., the first cut is made at 0.5 (50% of the samples goes to each model), the second cut close to 0.25 (25% goes to each model), the third close to 0.125, etc. On the other hand, with a high A, enabling the structure of the skewed Y to come into play, the shape of the dendrograms are changed more profoundly and somewhat unpredictably. In order to evaluate the predictive power of the set of PLS models building up each PLS-tree, we used a horizontal ruler placed at the Y-axis = 0.35. All branches existing at this cut-off and at higher values were evaluated for predictive power. Each local model was autofitted, and RMSEP was calculated for all 379 compounds in the test set and only for those compounds fitting the model. In eight out of the nine PLS-trees, it is possible to find a local PLS model enhancing the predictive power. Thus, there seems to exist a structure-activity theme that is invariant (or almost so) to the settings of A and B. With the most optimistic evaluation, RMSEP is lowered by approximately 40% thanks to the PLS-tree.

In order to accomplish training and prediction sets of compounds in the log KOC data set, the observations were first enumerated as 1,2,3,1,2,3,1... etc. This gave three sub-sets each comprising $351/3 = 117$ compounds. In the first round of calculations, sub-sets 1 and 2 were used as the training set and sub-set 3 as the prediction set. The roles of the three sub-sets were then fully permuted. Moreover, based on previous findings, the number of combinations of A and B was reduced from nine to five. The following five combinations were used: $A=0.1/B=0.1$; $A=0.5/B=0.1$; $A=0.3/B=0.3$; $A=0.1/B=0.5$; $A=0.5/B=0.5$. Thus 15 PLS-trees were computed. The resulting dendrograms are not plotted. The evaluation of the predictive power was done as above. The conclusion is that local modeling reduces RMSEP. This reduction is between 3% and 25% depending on which configuration of the training set is used. Additionally, the aim with this data set is to indicate which kind of interpretations can be carried out using the PLS-tree. To this end, we have chosen to zoom-in on a model called M14 (arising from training set 1&2 and $A=B=0.1$). Figure 5 shows the relationship between Y_{obs} and Y_{pred} for M14. By marking the compounds covered by M14 and referring to the same plot for the original reference model, we realize that M14 is focused towards the low-end part of the Y-variable.

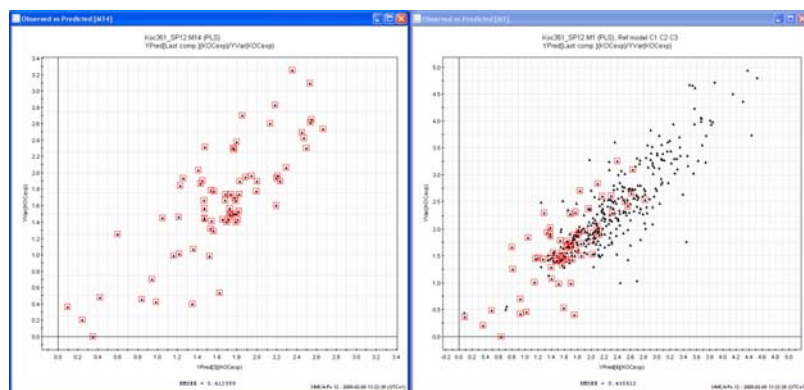


Figure 5: (left) Y_{obs} vs. Y_{pred} for M14. (right) Same for the reference model.

It is of interest to try to pinpoint if M14 has picked up some of the known classes. Figure 6 shows that this is indeed the case. The compounds of chemical class C are all allocated to this branch of the PLS-tree (recall that they are not all marked because a third of them were deliberately put in the prediction set). As opposed to this result, the compounds of

chemical class L are completely ignored by this part of the PLS-tree. Hence, the tentative interpretation is that the current PLS-tree is able to pick up known chemical classes. In that sense, the PLS-tree can be regarded as a confirmatory tool capable of corroborating clustering that is believed or determined to exist in a data set.

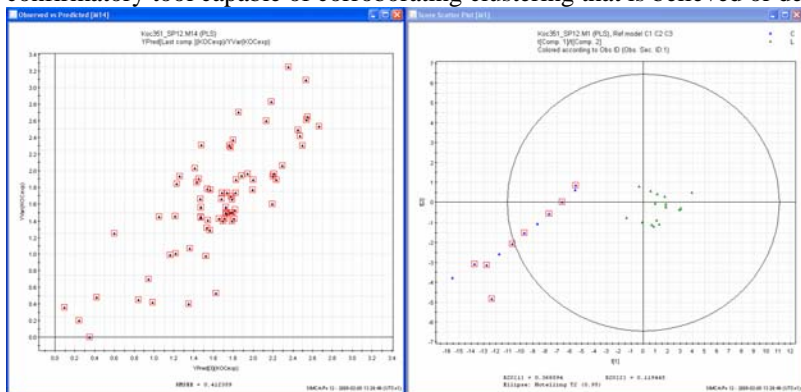


Figure 6. (left) Y_{obs} vs. Y_{pred} for M14. (right) Score space $t1/t2$ for the original reference model. Note that this plot only shows a sub-set of the observations earlier plotted in the right-hand part of Figure 3.

Model interpretation also includes the possibility of comparing loading related parameters between various models. Figure 7 shows scatter plots of the first loading vector for a case with similar and a case with dissimilar loading profiles. Models M14 and M27 (Figure 7, left), which encode similar RMSEP values, have similar loading profiles. This suggests that the QSAR themes identified by these models are reminiscent of one another. Conversely, models M14 and M261 (Figure 7, right), which also result in similar RMSEPs, but are based on different training sets, seem to have identified QSAR themes looking very different on closer examination.

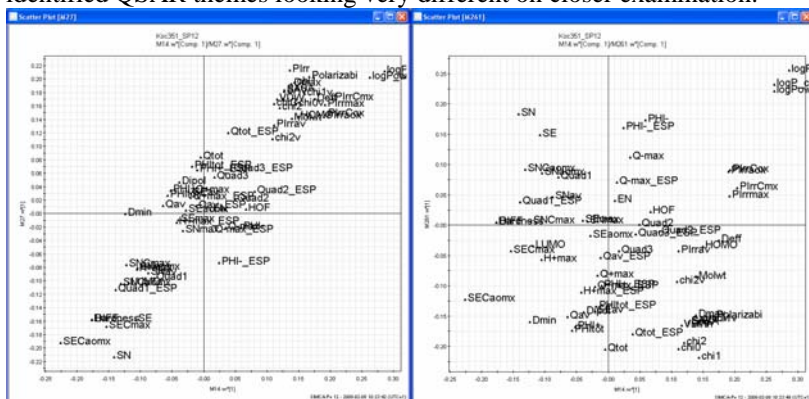


Figure 7. Scatter plot of the first PLS loading vectors between similar (left) and dissimilar (right) models.

5 Conclusion

PLS-trees, as implemented in SIMCA-P+™ 12, provide an interesting and rapid approach to data mining and clustering of large data sets. The PLS-tree approach can, like any PLS model, handle multiple and collinear variables, even more numerous than the number of observations. Moderate amounts of missing data are automatically handled by the PLS-tree algorithm. This is in sharp contrast to many classical clustering approaches that operate on some kind of distance matrix; these work well only up to around 10 K observations, and only with few variables.

The work reported here shows that the PLS-tree functionality may lead to:

- improved predictive ability due to local modeling;
- recognition/confirmation of clustering believed or anticipated to exist in data;
- enhanced interpretative options and understanding of local X/Y-themes in the data.

On a more general level, the conclusion is that fixing the parameters $A=B=0.3$, seems to produce good and representative results. For end-users wishing to quickly get into the usage of PLS-trees, these values are recommended.

6 References

- [1] Kriegl, J.M., Eriksson, L., Arnhold, T., Beck, B., Johansson, E., and Fox, T., Multivariate modeling of cytochrome P450 3A4 inhibition. *European Journal of Pharmaceutical Sciences* 24, 451-463, 2005.
- [2] Eriksson, L., Johansson, E., Müller, M., and Wold, S., On the selection of training set in environmental QSAR when compounds are clustered, *Journal of Chemometrics* 14, 599-616, 2000.