chem-bla-ics

QSPR modeling with signatures

Egon Willighagen

Published June 20, 2010

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/8d6w8-avm05

Keywords
Cdk, Chemometrics

Copyright
Copyright © Egon Willighagen 2010. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

I had to dig deep to find posts on QSAR modeling. There are quite a few on QSAR in Bioclipse, but that focuses on the descriptor calculation. In a quick scan, I could only spot two modeling posts:

- The CDK/Metabolomics/Chemometrics Unconference results
- When to stop including QSAR model variables...

Given the prominent place QSAR has in my thesis , this is somewhat surprising. Anyway, here is some more QSAR modeling talk.

Gilleain implemented the signature descriptors developed by Faulon et al. (see doi:10.1021/ci020345w; I mentioned the paper in 2006), and the CDK patch is currently being reviewed. With some transformations, the atomic signatures for a molecule can be transformed into a fixed-length numerical representation: `[70:1, 54:1, 23:1, 22:1, 9:9, 45:2]`. This string means that atomic signature 70 occurs once in this molecule and signature 9 occurs nine times. At this moment, I am not yet concerned about the actual signature, but just checking how well these signature can be used in QSPR modeling.

Rajarshi's fingerprint code provides a good template to parse this into a X matrix in R:

For my test case, I have used the boiling point data I used in my thesis paper *On the Use of 1H and 13C 1D NMR Spectra as QSPR Descriptors* (see doi:10.1021/ci050282s). Some of this data is actually available from ChemSpider, but I do not think I ever uploaded the boiling point data. This constitutes a data set with 277 molecules, and my paper provides some reference model quality statistics; that way, I have something to compare against. Moreover, I can use my previous scripts to do the PLS modeling (there are many tutorials online, but you can always buy an expensive book like the one shown on the right, if you really have to), (10-fold) cross-validation (CV), and 5 repeats of random sampling.

I strongly suggest people interested in statistical modeling to read this interesting post from Noel: whatever test set sampling method you use, you ***must*** do some repeats to learn about the sensitivity of your modeling approach to changes in the data set. Depending on the actual sampling approach, you might see different sizes of variance, but until you measure it, you will not know. For my application, these are the numbers:

```
# source("pls.R")
Read 277 items
   rank=66     LV=42     R2=0.987     Q2=0.921  RMSEP=31.37
   rank=66     LV=42     R2=0.983     Q2=0.924  RMSEP=12.405
   rank=65     LV=42     R2=0.985     Q2=0.949  RMSEP=38.503
   rank=63     LV=42     R2=0.983     Q2=0.948  RMSEP=36.981
   rank=65     LV=42     R2=0.986     Q2=0.923  RMSEP=21.49
   rank=64     LV=42     R2=0.983     Q2=0.91  RMSEP=17.759
   rank=64     LV=42     R2=0.983     Q2=0.921  RMSEP=17.062
   rank=66     LV=42     R2=0.986     Q2=0.94  RMSEP=40.311
```

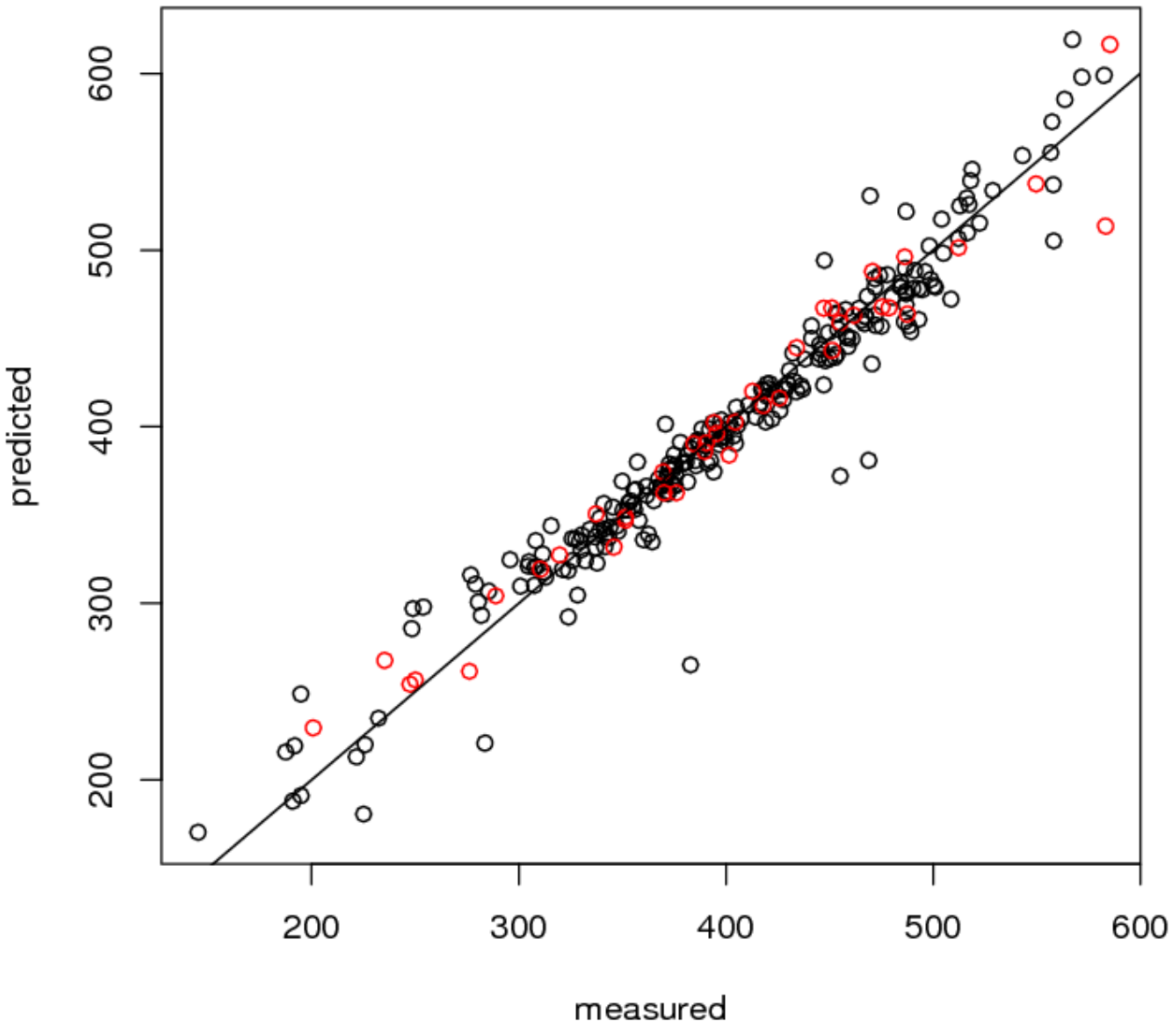
chem-bla-ics

```
rank=66     LV=42     R2=0.982     Q2=0.927  RMSEP=13
rank=68     LV=42     R2=0.986     Q2=0.929  RMSEP=16.23
```

I know 42 is the answer to the universe, but 42 latent variables (LVs)?!? Well, it's just a start. A more accurate number of LVs seems to be around 15, but my script had to make the transition from the old pls.pcr package to the newer pls package. And I have yet to discover how I can get the new package to return me the lowest number of LVs for which the CV statistic is no longer significantly different from the best (see my paper how that works). Actually, I have set the maximum LVs to consider to 1/5th of the number of objects (which is about the accepted ratio in the QSAR community); otherwise, it would have happily increased.

However, the five repeats nicely show the variance in the quality statistics, R^2, Q^2, and root mean square error of prediction (RMSEP). From the numbers, a model with $Q^2 = 0.94$ is **not** better than one with $Q^2 = 0.93$ (and I have seen the variance quite some larger). Bottom line: just measure that variability, and put it in the publication, will you??

Anyway, what we all have been waiting for: the prediction results visualized (in black the CV predictions; in red the test set predictions):



chem-bla-ics

Well, there is still much work to do, and you can expect the result to get better. Part of statistical modeling is to find the source of variance, and I have yet to explore a few of them. For example, what are the effects of:

- creating signature from the hydrogen-depleted graph
- effect of tautomerism (see this special issue)
- effect of the height of the signature

And there are so many other things I like to do. But this will do for now.