BERKELEY LAB

David Perryman <dperrym2@lbl.gov>

Prediction of Acute Aquatic Toxicity Toward Daphnia magna by using the GA-kNN Method

David Perryman <dperrym2@lbl.gov> Mon, Aug 10, 2020 at 4:40 PM
To: Davide Ballabio <davide.ballabio@unimib.it>

Davide,

Thank you very much for the clarification. I appreciate the fast response and clarity.

- Elliott

On Mon, Aug 10, 2020 at 3:09 PM Davide Ballabio <davide.ballabio@unimib.it> wrote:

> Dear Elliott,
> perfectly clear.
>
> 1) Q_ext^2 for external set: exactly, we used the Q2F3 formula, as
> described in the 2009 paper by Consonni (Comments on the Definition of
> the Q2 Parameter for QSAR Validation). We believe this is the best way
> (beside root mean squared error) to estimate regression on external
> molecules (see also
> https://pubs.acs.org/doi/abs/10.1021/acs.jcim.6b00277
> <https://pubs.acs.org/doi/abs/10.1021/acs.jcim.6b00277>), however this
> function can have drawbacks when comparing regression performances on
> external sets for models trained on different training sets (as
> described here:
> https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800029
> <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800029>).
>
> 2) cross validation Q2 was calculated with the formula (2), page 1670 of
> the Consonni 2009 paper. In fact, here we used internal five-fold
> cross-validation; this means training molecules were divided in 5 groups
> (I guess with a Venetian blind scheme). As an example, the split based
> on venetian blinds with e.g. 3 cross-validation groups would be for the
> first group [t,0,0,t,0,0,....,t,0,0], while the second one will be
> [0,t,0,0,t,0,....,0,t,0] and the third one [0,0,t,0,0,t,...,0,0,t],
> where t are the molecules included each time in the cross-validation
> groups. Thus, each molecule is left out just once and the formula at
> page 2 can be applied.
>
> Hope this helps.
>
> Best,
> Davide
>
>
>
> Il 10/08/2020 17:26, David Perryman ha scritto:
> > Hi Davide,
> >
> > I hope this email finds you well. I read your 2014 paper and have a
> > question about it.
> >
> > I understand how you do the coefficient of determination for training,
> > but I do not understand the coefficient of determination for cross
> > validation and for the test set. The paper you cite for the metric
> > Q_ext^2 (Consonni 2009) has 3 different functions: F1, F2, and F3. The
> > paper seems to hint that F3 may be best, and I am guessing that is

> what you used, but I want to be sure. Do you use the F3 function from
> Consonni 2009 for Q_ext^2 with the external set being 20% of the data
> randomly selected? And for cross validation, do you also use the F3
> function with the external data being the validation part of the five
> fold cross validation set?
>
> I hope this email is sufficiently clear about what I am asking. I
> liked your paper and found it clear and helpful, so thank you. If
> there is something I can do to be more clear about what I am asking,
> please let me know. Thank you for your time.
>
> - Elliott Perryman

--

Davide Ballabio,
Associate professor
Milano Chemometrics and QSAR Research Group
Department of Earth and Environmental Sciences
University of Milano-Bicocca - Italy
http://www.michem.unimib.it/

Tel: ++39 0264482818
mailto:davide.ballabio@unimib.it

--