Review of: "Identifying interactions in omics data for clinical biomarker discovery"

Antonio Parziale[1]

1 University of Salerno

Potential competing interests: The author(s) declared that no potential competing interests exist.

The authors applied the QLattice, symbolic regression machine learning algorithm, to some public datasets in the fields of proteomics, genomics, epigenomics and multi-omics. The aim of the authors is to show how interpretable methods are useful to extract more information from data and to understand dependencies among features and between features and the dependent variable.

My comments in detail:

- The adopted method (QLattice) must be commented by providing more information on how a model is built and how features are selected. Moreover, the set of operators and mathematical functions that could appear in a model should be listed and their semantics and syntax should be explained. Eventually, it sounds like the method allows to limit the maximum number of features that a model can contain but that's not reported in the method section.
- If I understand correctly the experimental protocol adopted by the authors, they split the data in training (80% of data) and test sets and they used the training set to generate the list of models and then select one of them. Then, selected one of the models, they applied a 5-fold cross-validation to measure the model performance. I checked the code provided as supplementary material and it seems that the protocol is the one I described before. This protocol is not correct, there is a bias because part of the data used as test data in the k-fold cross-validation was used as training data to build the model. The 5-fold cross-validation should be applied only on the test in order to find the best model and then measure its performance on the test set. Another solution could be avoiding the training/test partition but using only a K-fold cross-validation.
- For the epigenomics case, the authors should motivate/explain the reason behind a feature selection before the use of QLattice
- On page 8 the authors wrote, “As can be seen in Fig. 9, the primary separator of the two features”. I think they wanted to write “of the two classes” instead of the two features
- I think there is a mistake in the y-axis label in Figure 9, on the top-left graph. The y-axis should represent the probability, not the feature values.
- In Figure 10 it's not clear what are the two features highlighted in blue and from which model in Table 3 they are extracted. I suggest improving also the comment on page 8 related to figure 10

- What does “of the models at the head of each k-fold.” mean?