Review of dissertation thesis "Methods for Class Prediciton with High-Dimensional Gene Expression Data" by Jana Šilhavá

The study explores how to learn disease classifiers from data containing both clinical and gene-expression variables. This clearly stated problem is of obvious practical relevance and is also quite challenging from the theoretical point of view.

The thesis reads very smoothly. Relevant background is explained in a simple and straightforward way which indicates the author's good familiarity with the needed machine-learning and bioinformatics concepts. Also the language is almost flawless. My only complaint regarding the introductory parts are some statements over-generalized to the verge of falsity such as *"logistic regression cannot be used with with high dimensional data without a dimension reduction step or penalization"* or *"maximum likelihood estimates do not have a closed form."*

The key contribution of the thesis is really the exploration of the idea that a regression model is learned separately for each respective data source and then the linear forms pertaining to each of the data sources are summed prior to being transformed through the logit function. This is a neat idea because of its simplicity and because it allows to pull the right weapon on each different data source. In particular, the author applies vanilla logistic regression with the clinical data, and logistic regression with boosting with the microarray part of the data; this is called the LOG/Z+B/X method and it is compared to the separate baselines (i.e. LOG/Z and B/X, respectively). She also tries several extensions of LOG/Z+B/X, such as with a weight (tuned through validation) attached to each of the two summands. I particularly appreciated this latter extension since the weights in fact rectify the possibly incompatible scaling of coefficients of the LOG/Z and B/X submodels, respectively. As a sidenote, I think the author should have discussed this issue.

In the used benchmark data sets (a few breast cancer data sets and simulated data) the novel combined classifier usually performs better than classifiers learned from only one data source, which is reassuring.

However, there are a few points I disliked:

- In my opinion, the most important qualifier of the new method is how it performs in comparison to the most simple combination of the datasets where you simply put the clinical data as additional columns to the microarray data. I found no experiments answering this question.
- The second most important question is how the new method compares to other state-of-the art approaches which combine the data sources. The author acknowledges such approaches in Section 7.1 but never attempts to experimentally compare her method to them.
- The framework of the thesis allows many composed methods, for example,

in "B/Z+B/X", boosting would be applied to either data source. Why not try that, or for that matter, why not try also "B/(Z+X)" which would stand for boosting the simple method in the first item above. All in all, the choice of the tested methods seems too ad-hoc.

- Only a few real data sets are used for the experiments and thus the outcomes are not really conclusive. Since all the real data sets used in the thesis concern breast cancer, it would have been more appropriate to entitle the thesis accordingly.

- Different experiments are conducted with different subsets of the breast cancer data set library. For example, experiments reported in tables 7.8 and 7.9 are only conducted on two domains out of the five domains used for experiments reported in table 7.4. I could not find any explanation as to why this is the case. Again, the choices seem ad-hoc at best.

A further, minor concern is that tables 7.4 and 7.8 (and possibly other similar tables) are somewhat confusing in that the AUC of LOG/Z is reported under the $p = 50$ column which makes the reader wonder why it was *only* tried with $p = 50$ (of course, LOG/Z has nothing to do with p so it should be reported elsewhere).

Another claimed contribution is the study of the additional predictive value of microarrays in Chapter 8. Here I simply could not understand why the author presented again (though in a slightly different context) much of the material already presented in Chapter 7. In particular, the comparison between LOG/Z and LOG/Z+B/X readily follows from Chapter 7 and indeed, the AUC's reported in Table 8.1 are the same as reported in Chapter 7. Also, to my best understanding, the experimental scheme taking an entire page (pg 70) is exactly the same workflow as used in Chapter 7 for LOG/Z, and LOG/Z+B/X, respectively.

Chapter 9 contributes by showing that it is difficult to establish a few predictive genes valid over several studies, which emphasizes the need for the combined clinical-microarray approach. This point is a plausible, though not very novel, thesis. Minor glitches: Table 9.1 logically belongs into section 6.1 and should only be referenced in Chapter 9. The term *"predictive class prediction"* adopted in the text is a very silly name but it was not the author of the dissertation who had invented it.

The last contribution, presented in Chapter 10, concerns the exploitation of the gene ontology for feature (gene) selection. As the authors admits, the results are preliminary. I think this is also the case for the methodological design and the presentation. First of all, why does the author in Variant A add a value to the expression of a gene if that gene is present in a maximal clique in the semantic similarity graph? Not only the author provides no reason for such an operation but the operation makes no sense to me: I simply see no reason why, on the grounds of functional similarity of genes, one should pretend that the gene yields more mRNA than it does. Variant B literally repeats Variant A with slight alterations; would it thus not be better to say *"B is like A, except ..."*?

Variant C says how to compute a threshold but it does not say the main thing: what happens with the data in this variant? The decision to work only with 2000 *random genes* because *"the computations of SSMs were time consuming"* is not acceptable. If a random selection is neccessary then the entire experiment should at least be repeated several times and results averaged to reduce variance. However, why not e.g. pre-select 2000 genes using a fast (e.g. greedy) feature selection method and only then apply SSM? Also, why are the results reported on only one data set? In the chapter-specific conclusions (10.4) there is no mention whether the method helped or not, whether it performed good or bad, etc.

Considering the clearly stated and relevant research problem and the author's plausible and functional approach to solving it on one hand, and the relatively large number of raised issues on the other hand, this is a borderline thesis. **My final call is, however, that the pros outweigh the cons and the candidate should be granted the doctoral degree.**

In Prague, on Jan 22, 2013

Filip Železný
Czech Technical University in Prague