Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch 3 (2014) 78-82
ISSN 2295-3159

Using New Models to Analyze Complex Regularities of the World: Commentary on Musso et al. (2013)

Petri Nokelainen[a], Tomi Silander[b]

[a] University of Tampere, Finland

[b] Xerox Research Centre Europe, France

Abstract

This commentary to the recent article by Musso et al. (2013) discusses issues related to model fitting, comparison of classification accuracy of generative and discriminative models, and two (or more) cultures of data modeling. We start by questioning the extremely high classification accuracy with an empirical data from a complex domain. There is a risk that we model perfect nonsense perfectly. Our second concern is related to the relevance of comparing multilayer perceptron neural networks and linear discriminant analysis classification accuracy indices. We find this problematic, as it is like comparing apples and oranges. It would have been easier to interpret the model and the variable (group) importance's if the authors would have compared MLP to some discriminative classifier, such as group lasso logistic regression. Finally, we conclude our commentary with a discussion about the predictive properties of the adopted data modeling approach.

Keywords: Artificial Neural Networks; Commentary; Model-fit; Generative and Discriminative Models; Algorithmic Data Modeling

1. Introduction

Statistical methods are constantly developed, not only within statistics, but also in other disciplines, such as, physics, economics, bioinformatics, linguistics, and computer science. We therefore are very sympathetic to the attempts to promote new methods for analyzing educational data. However, also in this research field, for years there has been an emphasis on the predictive modeling, for example, to learn structures from the data (Nokelainen, Silander, Ruohotie, & Tirri, 2007; Tirri, Nokelainen, & Komulainen, 2013) and to predict class membership (Nokelainen & Ruohotie, 2009; Nokelainen, Tirri, Campbell, & Walberg, 2007; Villaverde,

Corresponding author: *Tomi Silander, Xerox Research Centre Europe, www.xrce.xerox.com, tomi.silander@xrce.xerox.com ; Petri Nokelainen, University of Tampere, www.uta.fi/edu, petri.nokelainen@uta.fi.*

http://dx.doi.org/10.14786/flr.v2i1.107

Godoy, & Amandi, 2006). The recent boom of data analytics has further increased the efforts in this front. One methodological rationale behind this development is that predictiveness guards against over-fitting and serves as a natural criterion for the quality of the model. Classical statistical literature was not emphasizing this aspect, since models were kept relatively simple to avoid over-fitting and to keep the calculations reasonable. In addition, much of the theory concerned asymptotic behavior in which case over-fitting is usually not an issue.

Increased computing power now allows more complicated models, such as Bayesian, fuzzy and neural networks, to be used. While the increased flexibility brings benefits, there are also possibilities to make new kind of errors in the analysis. Since we share the enthusiasm to promote new methods, we also feel that is of utmost importance to perform the analyses with these new methods using extremely high methodological standards. In this respect, we find some of the procedures followed in the recent article by Musso, Kyndt, Cascallar and Dochy (2013) problematic. Before discussing about these issues in detail, we wish to indicate that we agree with Edelsbrunner and Schneider's (2013) previous commentary on this article where they state that there are other data analysis techniques with similar properties than ANNs, but without the drawbacks.

2. Fitting to the test data

Our first concern is the reported 100% classification accuracy in such a complex domain, and the lack of thorough discussion of this issue. Multilayer perceptron (MLP) neural networks are universal function approximators (Lek & Guegan, 1999). With enough twisting of the parameters, one can use them to implement any classification rule (Schittenkopf, Deco, & Brauer, 1997). Consequently, the networks could in theory also be designed to explain the version of the dataset in which the GPA scores would be randomly assigned to the students. What knowledge does such a model (that can explain anything) extract from the real world?

One persuasive answer does indeed lie in prediction. Only the regularities help one to generalize beyond the training sample, that is, to predict. But here one needs to be very careful. To do this right, the data must first be split into two parts and then the model must be built using only the first part. The testing should be done with the second part of the data – the part that was not used in the model building process at all. The big question is: Can we trust us to be able to refrain from "cheating" (using the test data)? In order to avoid this, it would be best to gather the test data after building the model, or to separate it from the training data in the very beginning, and give it to somebody else who will then, after the model has been built, test the accuracy of the model – once and for all!

The paper by Musso and her colleagues (2013) practically acknowledges that such a discipline was not rigorously followed. The network structure and learning parameters were adjusted to maximize the accuracy in test data. Many models were tested to achieve this. Even the division of the data into training and test samples was manipulated in order to "… maximize the training sample while preserving the appearance of all detected patterns in the testing sample …" (Musso et al., 2013, 60).

Now, one cannot totally exclude the possibility that the authors actually promote this methodology as a sound one. Take a maximally flexible model family, find the most parsimonious model that fits 100% to the data, and then analyze the model. But if that were the case, why torture oneself with the tedious manual work to find 100% fit (yes fit, not generalization) to the test data? It would be easier to just fit to the whole data set – but that would break the illusion of prediction.

3. Comparison with the linear discriminant analysis

We find that the authors' decision to compare the model to the other models sets a very good example that should more often be followed in the educational research. Such comparisons are widely used in machine learning (e.g., Demšar, 2006). However, comparing the multilayer perceptron and the discriminant analysis raises some questions. Behind the linear discriminant analysis is a linear discriminant model that defines a joint probability distribution for the whole 19-variate (18 independent variables + GPA

class) data vector. Such joint probability distributions can be used for classification, since the conditional probability P(GPA-class | predictors) is proportional to the joint distribution P(GPA-class & predictors). These kinds of classifiers are usually called generative classifiers (e.g., Xue & Titterington, 2008), since they are based on the models that can be used to sample the whole (19-variate) data vectors.

MLPs are not generative classifiers, but so called discriminative classifiers. They are built to directly estimate the conditional distribution P(GPA-class | predictors) without modeling the relationships among the predictors. (Reading the paper sometimes makes you feel that the authors claim otherwise.) While the linear discriminant model DA1 used in the Musso et al. (2013) paper has about 2*18 + 18*18 = 360 parameters, the neural network model has 18*15*2 = 540 parameters. The difference in number of parameters is not huge, but all the parameters of the MLP are used for modeling the conditional distribution, while the parameters in the linear discriminant model also take care of modeling the relationships between variables.

Since the linear discriminant is also a predictive classifier, one cannot but wonder why the confusion matrices for linear discriminants were not reported. Those numbers surely would have fitted to the same space without any problem. On the other hand, it is plausible that any differences found are due to the other classifier being generative and the other one discriminative. It would have been much more meaningful to compare the MLP to some discriminative classifier such as a logistic regression, or better yet, some sparse version of it such as the group lasso with interaction terms (Meier, van de Geer, & Bühlmann, 2008) that would make interpreting the model and the variable (group) importances much easier. Furthermore, the Musso et al. (2013) paper is very unclear about how the variable importances have been calculated. The attempt to follow the references only lead to the statements like “this has been implemented in software X” or to an unpublished technical report by one of the authors.

4. Two cultures

According to Breiman (2001b), there are two statistical modeling cultures. The Data Modeling Culture assumes that the data are generated by a given stochastic data model (such as linear or logistic regression). The Algorithmic Modeling Culture treats the data mechanism as unknown, using, for example, decision trees and neural networks. Although the first of these two cultures, focusing on data models, is still dominating, many fields outside statistics are rapidly adopting a wide variety of tools.

Neural networks are often considered as black-box models that do not offer a good explanation and understanding of the domain (Correa, Bielza, & Pamies-Teixeira, 2009). Consequently, such models are sometimes hastily deemed as unsuitable for much of the science. We would like to take the opportunity to say a word for such black-box models along the lines expressed by a statistician, Leo Breiman (1928-2005). World may not be a simple place. While among the simple theories there are those who most closely approximate the complex reality, it is a priori possible that none of those simple theories, even the best of them, approximate the situation well. If the model does not predict well, one can argue that it has not captured the regularities of the world, so what insight would understanding and interpreting such a model offer us. (Breiman, 2001b.) Most of the statistical community would agree that only if the model is reasonably good (and we mean generalization, not just fit to the sample), interpretation makes sense.

Edelsbrunner and Schneider (2013) indicate in their commentary on this article that whenever possible, more theory-driven data modeling techniques should be preferred. However, if we limit ourselves to the models that can be easily interpreted, we may end up discarding models that truly capture important regularities of the domain.

There are two different strategies then to extract true knowledge from the world. The first one is a classical one in which we try to find a well predicting model among the easily interpretable ones. This is the path that should always be attempted. Unfortunately, we suspect that it was not seriously pursued in the article by Musso et al. (2013). It is also possible to try to build a well predicting model, even if it is not that easy to interpret, and then put more effort to squeeze out the knowledge from the model. One could argue that this is what happens, when you ask a doctor why she made the diagnosis she did. The answer will (only) be some approximation of the real reason. Still doctors are considered useful.

We have an educated guess, that such a procedure is behind the independent variable importance measures featured in the article. Naturally, such procedures should be carefully documented in order to understand what kind of information we have managed to extract from the model. The article leaves the impression of the claim that artificial neural networks were somehow especially good for inferring how different complex patterns of variables affect the outcome. However, the presented results list only univariate importance of variables. How could that possibly tell us anything relevant about complex patterns?

Neural networks are by no means the only black-box models that can be successful in the prediction. Many ensemble learning based or motivated methods, for instance random decision forests (Breiman, 2001a) and Bayesian additive regression trees (Chipman, George, & McCulloch, 2010), are among such models. Ensemble methods have reached very high classification accuracies by using several (or growing a forest of) decision trees on the same data instead of a single-tree predictor.

Ever increasing data sizes (e.g., Massive Open Online Courses, MOOCs, may have 100 000 students with all their data gathered automatically to the digital form) and increasing computer power may well shift focus from small, simple and understandable models, to the big, complex black-box models. But hopefully some of that computing power can also be used to extract understandable (even if not always very close to truth) approximations of the true complex regularities of the world.

Key points

- Artificial neural networks (ANN) certainly provide interesting modeling possibilities for educational scientists, but they also set certain challenges for the design of the study and interpretation of the results.
- The article shows a very good example by comparing the results of ANN to a conventional data modeling approach, but the comparison should have been made between two discriminative classifiers.
- Ensemble methods provide a modern and powerful alternative to neural networks as they use predictions of several models built during learning process instead of using a single model.

References

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*, 5–32. doi:10.1023/a:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231. doi:10.1214/ss/1009213726

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, *4*(1), 266–298. doi:10.1214/09-aoas285

Demšar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Correa, M., Bielza, C., & Pamies-Teixeira, J. (2009). Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. *Expert Systems with Applications*, *36*, 7270–7279. doi:10.1016/j.eswa.2008.09.024

Lek, S., & Guegan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*, 65–73. doi:10.1016/s0304-3800(99)00092-7

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, *70*(Part 1), 53-71. doi:10.1111/j.1467-9868.2007.00627.x

Musso, M. F., Kyndt, E., Cascallar, E. C., & Dochy, F. (2013). Predicting general academic performance and identifying differential contribution of participating variables using artificial neural networks. *Frontline Learning Research*, 1, 42-71. doi:10.14786/flr.v1i1.13

Nokelainen, P., Silander, T., Ruohotie, P., & Tirri, H. (2007). Investigating the Number of Non-linear and Multi-modal Relationships between Observed Variables Measuring a Growth-oriented Atmosphere. *Quality & Quantity*, *41*(6), 869-890. doi:10.1007/s11135-006-9030-x

Nokelainen, P., & Ruohotie, P. (2009). Non-linear Modeling of Growth Prerequisites in a Finnish Polytechnic Institution of Higher Education. *Journal of Workplace Learning*, *21*(1), 36-57. doi:10.1108/13665620910924907

Nokelainen, P., Tirri, K., Campbell, J. R., & Walberg, H. (2007). Factors that Contribute or Hinder Academic Productivity: Comparing two groups of most and least successful Olympians. *Educational Research and Evaluation*, *13*(6), 483-500. doi:10.1080/13803610701785931

Schittenkopf, C., Deco, G., & Brauer, W. (1997). Two Strategies to Avoid Overfitting in Feedforward Networks. *Neural Networks*, *10*(3), 505-516. doi:10.1016/s0893-6080(96)00086-x

Schneider, M., & Edelsbrunner, P. (2013). Modelling for Prediction vs. Modelling for Understanding: Commentary on Musso et al. (2013). *Frontline Learning Research*, *1*(2), 99-101. doi:10.14786/flr.v1i2.74

Tirri, K., Nokelainen, P., & Komulainen, E. (2013). Multiple Intelligences: Can they be measured? *Psychological Test and Assessment Modeling*, *55*(4), 438-461. doi:10.1007/978-94-6091-758-5_1

Villaverde, J. E., Godoy, D., & Amandi, A. (2006). Learning styles' recognition in e-learning environments with feed-forward neural networks. *Journal of Computer Assisted Learning*, *22*, 197–206. doi:10.1111/j.1365-2729.2006.00169.x

Xue, J-H., & Titterington, D. M. (2008). Comment on "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes". *Neural Processing Letters*, *28*(3), 169-187. doi:10.1007/s11063-008-9088-7