Model Validation in Biological Applications of Evolutionary Computation

Marylyn D Ritchie

Center for Human Genetics Research, Department of Molecular Physiology and Biophysics,
Vanderbilt University, 519 Light Hall, Nashville, TN 37232
{ritchie@chgr.mc.vanderbilt.edu}

Abstract. Model validation has become a strong component of successful statistical analyses. Some researchers in the evolutionary computation (EC) community have applied various model validation strategies. However, in general, model validation is underused in biological applications of EC. This essay describes the importance of model validation and some techniques that have been performed. Unfortunately, there is not an optimal method for all data types. Thus, this essay is not meant to review all possible methods and describe the best one for EC applications. Instead, the goal is to introduce the concepts in hopes to spark further research and discussion of this important topic.

1 Introduction

Bioinformatics, or the union of computer science and biology, has forever changed our ability to explore the complexities of biological data. Advancing technology in the laboratory has provided an enormous explosion of information. Without bioinformatics, the interpretation of all this data would be impossible. Identifying statistical and computational models that characterize biological processes is one of the main goals of bioinformatics. The approaches used to achieve this range from traditional statistical analysis, to novel statistical methods, as well as artificial intelligence and machine learning. In recent years, evolutionary computation (EC) has emerged as an important field for bioinformatics. Since EC has adopted many of its primary features from biology, it is time for EC to return the favor by aiding in the daunting task of understanding a wealth of biological phenomena.

EC has been applied to biological data in many capacities including genetic algorithms (GA), genetic programming (GP), grammatical evolution (GE), artificial immune systems, etc. In addition, the biological applications have been from numerous areas of biology including gene expression analysis, sequence alignment, proteomics, and protein structure prediction to name a few. Despite the biological problem or the EC methodology utilized, these previous applications share a common goal: to identify one or more models that classify or predict biological data.

While EC is a powerful approach for exploring large search spaces, there are several challenges when using EC methods for biological data analysis. First, depending on the question at hand, the resulting model from an EC search may be very difficult to interpret. This challenge is not unique to EC methodologies. Many statistical and computational methods, while powerful in detecting biological models, are un-interpretable to biologists. This is an area that will require further research. Second, many biological problems are highly dimensional and have an effectively infinite search space (in terms of exhaustive search capabilities). While EC methods are better equipped for exploring the vast territory related to biological problems, they are very computationally intensive and can still miss certain features of the search space. This challenge can be met through the use of parallel supercomputing technologies as well as optimal EC parameter settings to avoid stalling on local optima and premature convergence. Finally, due to its stochastic nature, each run of an EC method may produce a different best model composed of different variables. This is a problem when the goal of a study is to identify the best plausible model of some biological phenomena, such that further "wet" lab research can validate the findings. If each run of a GP or GA yields a different model, which model should be reported as best? In addition, if the goal is to find a more general predictive model, how can we prevent the EC algorithm from overfitting the sample data set?

Model validation is a technique that may be able to rectify the model selection issue as well as the overfitting concern. Statistical modeling using traditional methods such as linear discriminant analysis and logistic regression suffer from these same problems. The importance of model validation in these traditional techniques has become a more recognized component of sophisticated statistical analyses. Here in the EC field, some researchers are embracing model validation procedures. However, the importance of such validation demonstrates that model validation should become a part of all real data applications with the exception of performing classification of a single data set to yield the smallest possible error rate.

A caveat of suggesting/urging the use of model validation procedures is that there is not one optimal method to follow. Several model validation techniques exist, each having its own strengths and weaknesses. In this essay, some of these model validation techniques are introduced including the benefits and drawbacks of each approach.

2 Methods

Cross validation is a statistical technique where a model is developed on a subset of the original data and tested or validated on a portion of the data that was not used in model building. The ultimate validation of a model is to evaluate the predictive ability in a second data set, which is completely independent of the first data set. In reality, this is not typically feasible due to the expense of data collection. Instead cross validation is a good surrogate to test the model on unseen data.

The way that the data are split can vary in many ways from leave-one-out cross validation (LOOCV) where only one individual is left out of each analysis, to five-fold or ten-fold cross validation where the data are divided into five or ten partitions respectively and the model build on 4/5 or 9/10 and tested on 1/5 or 1/10 [1]. The analysis is performed on each possible split of the data such that all individuals are used in the test set only once. An additional variation on the "N" fold cross validation includes performing the data split multiple times (such as ten) and averaging the results across the splits [2]. Each type of cross validation has certain benefits. LOOCV provides an unbiased estimate of the prediction error. However, it has a high variance due to the similarity in the training sets [1]. Five-fold or ten-fold cross validation, on the other hand, has a smaller variance, but the estimate of the prediction error may be biased [1]. Finally, ten-fold cross validation ten times has the smallest variance and bias, but it is the most computationally intensive [2]. Simulation studies have been done to evaluate the "best" cross validation technique [2]. However, each data split has advantages and disadvantages and it is a trade-off based on what is most important for each specific study.

Depending on the fitness function used for selecting the best model, an additional type of cross validation may be even more robust to overfitting. N-fold cross validation typically uses the error of the training set as the fitness function. This value is likely to continue decreasing until it over fits. This may result in a model that classifies well, but does not generalize for independent data. A potential solution to this problem involves a three-way data split where one creates a training set, testing set, and validation set [3]. Here, a model is developed on the training set and evaluated on the testing set. Next, the best model is identified as the model where the training error and testing error have the absolute minimum difference. This model can then be validated on the final "validation" set. This technique may prove to be even more robust to overfitting in comparison to N-fold cross validation. However, the original data set must be large enough to maintain power while splitting into three subsets.

Finally, bootstrapping is another model validation strategy than can be explored [2]. Bootstrapping involves performing random sampling with replacement to create "X" new data sets that are the same size as the original data. The analysis is conducted on each bootstrapped sample to create a "best" model. This model is then tested on the original data. Some simulations suggest that bootstrapping may have smaller variance and bias than any cross validation approach. It is, however, substantially more computationally expensive. As mentioned earlier, there is not one optimal model validation procedure that generalizes to all classes of problems. Here a few possible strategies were mentioned, but others should be explored as well.

3 Conclusion

Model validation is not new to biological applications in evolutionary computation (BioGEC). Moore et al. [4] describe a symbolic discriminant analysis (SDA) technique

using LOOCV for gene expression analysis. In addition, Moore et al. developed a statistic to select the best model (variables) when multiple models are detected in the cross validation splits [5] This metric, cross validation consistency (CVC), has been used with other data analysis methods in addition to EC [6]. This statistic is a measure of how often the same variables occur in the models, divided by the total number of best models. Rowland described an application of GP to spectroscopy data using the three-way data split approach. Ritchie et al. [7] demonstrated a GP for optimizing neural network architecture using ten-fold cross validation and CVC for model validation and variable selection. These are only a few of the recent publications demonstrating model validation in BioGEC.

The goal of this essay was not to review all previous application of model validation techniques in BioGEC. Nor was it to suggest an optimal strategy for model selection and model validation. Instead, the goal was to explain some model validation procedures and give a few examples in EC applications to facilitate discussions between computer scientists, biologists, and biostatisticians such that researchers from these fields can work together to bring model validation to the forefront of all biological applications of evolutionary computation. It is my opinion that this area is underused and more research emphasis should be concentrated on selecting and validating EC models. If the goal is to learn something about biology, we need to use EC to find a solution that generalizes to the field of biology rather than a single data set.

4 Acknowledgements

This work was supported by institutional developmental funds from Vanderbilt University Center for Human Genetics Research.

5 References

1. Hastie T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, Berlin (2001)
2. Braga-Neto, U., Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? Bioinformatics 20 (2004) 374-380
3. Rowland J.J. Generalisation and model selection in supervised learning with evolutionary computation. In: Lecture Notes in Computer Science Vol 2611 ed. by: Raidl, G, et al. Springer-Verlag, Berlin (2003) 119-130
4. Moore J.H., Parker J.S., Olsen N.J., Aune T.S.: Symbolic discriminant analysis of microarray data in autoimmune disease. Genet Epidemiol 23 (2002) 57-69

5. Moore J.H.: Cross validation consistency for the assessment of genetic programming results in microarray studies. In: Lecture Notes in Computer Science Vol 2611 ed. by: Raidl, G, et al. Springer-Verlag, Berlin (2003) 99-106
6. Ritchie M.D., Hahn, L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F., Moore J.H.: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69 (2001) 138-147
7. Ritchie M.D., White B.C., Parker J.S., Hahn L.W., Moore J.H.: Optimization of neural network architecture using genetic programming improves detection of gene-gene interactions in studies of human diseases. BMC Bioinformatics, 4 (2003) 28