Data Fusion by Kernel Combination for Behavioral Data

D. Fekas[1], P. Zimmerman[2], C. J. F. ter Braak[3]

[1,3]*Biometris, Wageningen University, Wageningen, The Netherlands.* [1]*d.fekas@noldus.nl*

[2]*Noldus Information Technology bv, Wageningen, The Netherlands*

Introduction

Technological advances have radically changed the collection of behavioral data in animals. Modern tools complement, or even replace human observations in some labs. Also, behavior manifests itself in many different and complex ways, so behavioral scientists have developed a strategy of subjecting the animals to multiple observation systems, for longer periods of time, in order to get as much information as possible. Methodologically, the sample size (i.e. the number of animals) should be balanced from a statistical and an animal welfare point of view. Efficient statistical methods are therefore required to combine and analyze data from different types of experiments, while using a minimum number of animals. In this article, we present a generic framework that enables such a combination of diverse behavioral data.

Data description

We used behavioral observations of 4 strains (i.e. classes) of mice; a control group and three mutants which were designed to model Parkinson's disease (alpha-synuclein, synphilin-1 and double transgenic). There were 12 animals per strain and the raw data consisted of both video tracking data in a home cage (PhenoTyper®) and gait data from a gait analysis system (CatWalk® XT).

Video tracking within the home cage, which is performed by EthoVision® XT, produces large amounts of data, as the animals are tracked at a sample rate of 15 per second, during period of one week. There are predefined zones inside the cage and, also, automated cognitive tests take place at specific times. The output parameters are of a spatio-temporal nature: the 2D-body position in the cage, velocity, angular movement, relative position (within the cage) and shape are calculated on each frame.

The traditional way to analyze such longitudinal data is to split them in time-windows (bins) and then summarize (using mean values and coefficients of variation) all the calculated parameters per bin. We did this for both 12 hour long bins, corresponding to the light/dark cycle in the home cage, and for 1 hour long bins, taking only the first hour of the dark period for each day. However useful, the shortcoming of this approach is that if the bin is too long then you average over a very long period and hence lose information, while if too short, synchronization problems might appear; different animals might be engaged in different activities in each bin. Therefore, one could argue that comparisons across animals are not meaningful for short bins.

In this paper, we introduce an additional approach of representing behavior through strings. We define a set of behaviors of interest (alphabet) according to specific rules and then we produce strings or "words" that correspond to sequences of these behaviors in specific time periods. Consequently, we try to overcome the problem of misalignment in time by defining an appropriate similarity measure that detects similarities also between non-contiguous strings. This second approach results also in longitudinal, but very different type of

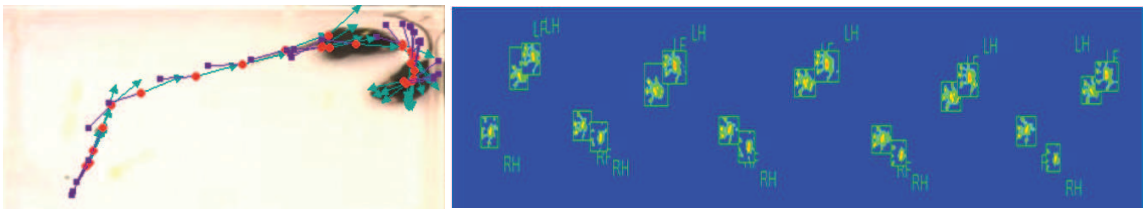

Figure 1. Snapshots of automated animal tracking (left) and gate analysis (right).

data; effectively a set of strings for each animal, whereas the time bin data sets are numeric. In addition, the string comparison method enables us to use the zone transitions (within the home cage) in the same classification context.

The gait analysis system on the other hand, produces numeric, but not particularly longitudinal data. CatWalk XT generates a large number of parameters related to a) individual paw prints (e.g., size of the print, mean intensity, stand, swing), b) spatial relationship between paws (e.g., base of support, print positions) and c) temporal relationships between paws (e.g., phase dispersion). All parameters are calculated for each analyzed run, i.e., a crossing from one side of the glass plate to the other on the CatWalk system. Runs can be acquired at different time points throughout the animal's life.

In total, from the video tracking data and gait analysis data we derived a total of 7 data sets, each measuring different aspects of the recorded behavior:

1. Bin12 - Using 12-hour long bins from the home cage, where we use means and coefficients of variation for each parameter as our input variables.
2. Bin1 - As above, but only using data from the 1st hour of the dark period each day.
3. String12 - 12-hour long string representations (from the home cage).
4. String1 - 1st-hour string representations (from the home cage).
5. Zone12 - Zone transition string (12-hour bin).
6. Zone1 - Zone transition string (1st-hour bin).
7. CatWalk - Gait analysis data.

We then tried to combine all this information sources in a single analysis.

Methods

In this research we tried to create a good classifier; a way to predict the strain of an animal, given the behavioral data from all the available information sources. That requires a model to be trained to find out what are the characteristic parameters for each strain. Therefore, even if it seems counter-intuitive at first glance to try to predict something we already know (the strain of each mouse), the process of doing so reveals a lot concerning the differences between groups.

A kernel is a similarity measure; when given a set of input variables and their corresponding classes, we can represent the relations between them using the kernel trick [1, 2]. For data fusion, we used a kernel combination approach [3, 4, 5, 6, 7]. The advantage of this method is that we can create one composite kernel for all the data, thus building one single classifier. The alternative would be to build a classifier for each separate data set, but then we would have to optimally combine them.

In short, the method works as follows. Consider a classification problem with Ş different feature spaces, from which we have D_s-dimensional predictor variables x_n^s, where s = 1,…, S and the corresponding target variables t_n ∈ 1,…, C for n = 1,…, N, with C the number of classes and N the number of observations. Our approach lies in embedding the information from the different feature spaces into Hilbert spaces using the kernel trick, and then combining these into one composite kernel space:

$$K^{\beta\Theta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{S} \beta_s K^{s\theta_s}(\mathbf{x}_i^s, \mathbf{x}_j^s),$$

where β is an S x 1 vector of weights and Θ is an Ş x D^s matrix containing the kernel parameters θ_s of the base kernels. Each of the base kernels Ķ^s is constructed based on our prior knowledge about the data source and its desired similarity metric. Unknowns are estimated as in [8].

Results

The classification performance can be measured by the error rate (i.e. the percentage of misclassified samples) using cross-validation [9]. Part of the data is used to train the classifier which we then use to predict the strain for the "left out" data.

In Figure 2 below, one can see the improvement in performance of the combined data (all) compared to the classification achieved by the individual data sets. It needs to be mentioned that this is a four-class problem which explains the big variation in the estimates. Wild types are much easier distinguished from the transgenic groups than the transgenic groups amongst each other.

Conclusions

In this article we provide a way to combine behavioral data from different tests in a classification framework and we present the improved classification performance using this approach. The data sets used in this experiment are used only to present the new possibilities. The main advantage of this approach is its flexibility. In essence, one can combine any kind of data from the same subjects in the same analysis. At the same time, we propose a novel behavior representation and analysis which renders it possible to compare behaviors that are not necessarily synchronized.

Ethical statement

The aforementioned data resulted from an experiment that was approved by the Ethical Committee of Utrecht University.

Acknowledgements

Dimitris Fekas is funded by the European Commission in the framework of Neuromodel, a Marie Curie Initial Training Network (ITN).

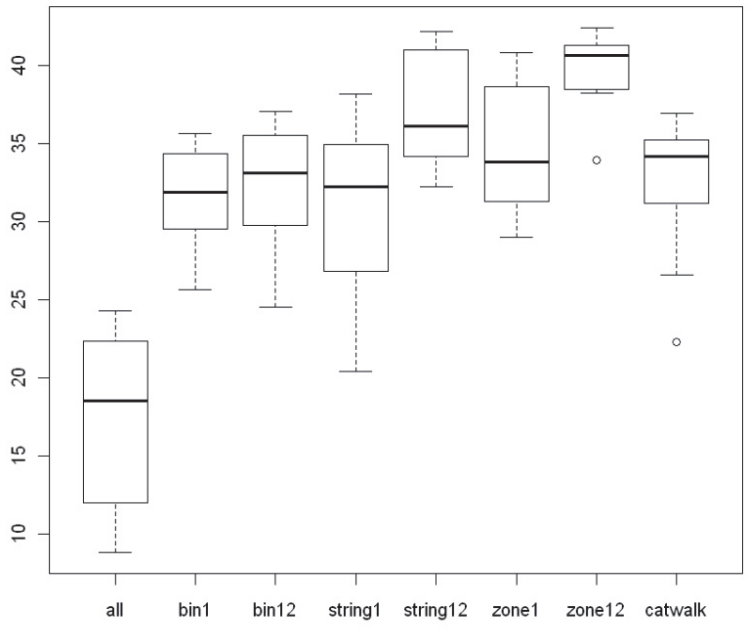

Figure 2. Error rate for the 4 class problem. The estimate of the combined data (all) is much lower than those obtained by each individual data set.

References

1. Shawe-Taylor, J., Cristianini, N., (2004). *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

2. Scholkopf, B., Smola, A.. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

3. Damoulas, T., Girolami, M. (2009). Combining feature spaces for classification. *Pattern Recognition* **42**(11), 2671-2683.

4. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research* **5**, 27-72.

5. Ong, C. Smola, A. Williamson, B. (2005). Learning the kernel with hyperkernels. *Journal of Machine Learning Research* **6**, 1045-1071.

6. Lewis, D. Jebara, T. Noble, W. (2006). Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics* **22**(22), 2753-2760.

7. Lee, W. Verzakov, S. Duin, R. (2007). Kernel combination versus classifier combination. *Multiple Classifier Systems (Proc. 7th Int. Workshop, MCS 2007, Prague, May 23-25, 2007), no. 4472*: 22-31.

8. Damoulas, T. Girolami, M. (2009). Pattern recognition with a bayesian kernel combination machine. *Pattern Recognition Letters* **30**(1), 46-54.

9. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, *Series B* **36**, 111-147.