Machine Learning Techniques for Parameter Selection and Automated Behavioral Classification

Dimitris Fekas
Noldus Information Technology
Nieuwe Kanaal 5,
6709 PA Wageningen,
The Netherlands
dimitris@noldus.nl

Raymond. C. de Heer
raymond.deheer@
deltaphenomics.com

Marco Mellace
marco.mellace@
deltaphenomics.com
Delta Phenomics BV
PO Box 80086, 3508 TB Utrecht,
The Netherlands

Berry M. Spruijt
Delta Phenomics BV,
P.O. Box 80086, 3508 TB Utrecht,
The Netherlands
berry.spruijt@deltaphenomics.com

Cajo J.F. ter Braak
Biometris, Wageningen University and Research Centre, Box 100, 6700 AC Wageningen,
The Netherlands
cajo.terbraak@wur.nl

ABSTRACT
In this paper, we address two different problems: efficient parameter selection for complex data sets and automated behavioral classification. Different strains of rodents were monitored in the PhenoTyper and various observable features have been quantified. We applied several machine learning techniques in order to explore the structure of the data and we managed to identify the most important behavioral parameters. In addition, we compared different classification techniques with respect to their accuracy and robustness in determining behavioral differences between the strains.

Author Keywords
Variable selection, behavioral classification, support vector machines, random forests, multidimensional scaling, multivariate analysis.

INTRODUCTION
In the area of behavioral research, there is a great need for standardized, automated experimental set-ups which allow for repeatable experiments and more reliable results. In order to tackle these issues, Noldus Information Technology has developed both hardware (PhenoTyper) and software (EthoVision XT) that allow automated video tracking of animals. Such an approach greatly improves our ability to quantify observable behaviors. However, the problem is deriving meaningful interpretations of the vast amount of data acquired by the tracking system. In this research, we compare the behavior of two different strains of mice with respect to various locomotive parameters.

MATERIALS
The PhenoTyper is a specially designed cage which allows automated observation of mice or rats. On top of each cage there is a unit containing the hardware needed for video tracking. In addition, other hardware devices are available, such as computer-controlled lights, a sensor to detect when the animal is drinking and a pellet dispenser. EthoVision XT is the video tracking software which detects the animal in the PhenoTyper. It also provides the option to the user to determine zones inside the cage according to the experiment and a Trial and Hardware Control module to control external devices such as lights, sounds and the pellet dispenser. EthoVision XT continuously samples the coordinates of the center of gravity of the animal from which the trajectory and velocity of the animal can be calculated. By using Trial and Hardware Control, it is possible to automate cognitive tasks and further discover behavioral characteristics such as anxiety, learning capability and memory. These tools allow for continuous recordings without much human interventions, and are used by behavioral researchers worldwide [6,9,12,13].

The animals used in this experiment were obtained by Harlan laboratories (males, 8-weeks-old): They were of 2 different strains: C57BL/6JOlaHsd and DBA/2JOlaHsd. The experiment was conducted at the research facilities of Delta Phenomics and it lasted for a period of one week. It

was initially designed to validate automated behavioral tests, which are not directly within the scope of the presented research, because our main focus is on locomotive behavior.

METHODS

For our analysis, a choice needed to be made for selecting a biologically meaningful set of parameters, as input for the classification algorithms. The raw EthoVision XT data consists of almost continuous measurements at a rate of 12.5 samples per second. Missing samples were interpolated and the entire set was smoothed using the LOWESS algorithm [5]. From the smoothed data, we can directly calculate some parameters for each animal, such as: distance moved, velocity, measures of the animal's elongation, mobility and angular velocity, duration and frequency of stops, as well as of visits to certain predefined zones within the PhenoTyper. During the experiment, a light/dark phase of 12/12 hours was used. Therefore, the above-mentioned parameters were either averaged or summed up for that time window. For the "continuous" parameters that were averaged, apart from the mean, the coefficient of variation was also used to provide some more information about their distribution in that rather long time bin.

For the final data set used, we collected these summary statistics for each of the parameters, each of the time bins and each of the animals and thus, we created a summary file for the entire experiment. After some standard preprocessing, such as removal of highly correlated parameters, logarithmic transformations (where appropriate) and scaling, we explored how the data was organized and we visualized that using methods like principal components analysis [7] and (non-metric) multidimensional scaling [3].

The main goal of this research was to discover which parameters mostly contribute to the between-strain differences in behavior. For variable selection, we used two methods: logistic regression [15] and support vector machine [10]. The results of both methods will be compared and behavioral implications of these parameters will be discussed. The second goal of this paper is to build a classifier, which can assign each animal to its respective strain, using behavioral parameters. With respect to this, our research continues as follows. Some of the animals are treated as if we did not know to which group/strain they belong to. The rest of the data is used as a training set and different algorithms that try to "learn" the machine how do animals from each strain behave. The classifier's performance can be measured by checking whether the animals that have been excluded from the training set are assigned correctly to their respective classes. The classification algorithms that were used were: random forests [2], support vector machines [11,14], and partially squared discriminant analysis [1]. The performances of the different classifiers were compared with respect to their accuracy and robustness using cross-validation.

CONCLUSIONS AND FUTURE WORK

The methods techniques mentioned above, are nowadays applied in numerous diverse scientific fields, but their use within behavioral science is to our knowledge quite limited [8] and hence there is much scope for novel developments. This is an on-going research and even though the importance of many behavioral parameters has been discovered, it may well be the case that there is more information in the data. For instance, some particular time bins may be more important than others and valuable information is lost when using a 12 hour time bin. Also, possible day-to-day differences could perhaps be explained by the cognitive tests that take place. This has not yet been taken into account, but it will be done in the near future.

ETHICAL STATEMENT

The aforementioned experiment was conducted after getting approved by the DEC of the University of Utrecht (DEC number: 05111201).

REFERENCES

1. Barker, M. and Rayens, W., Partial least squares for discrimination. *Journal of Chemometrics, 17* (2003), 166-173.
2. Breiman, L., Random forests. *Machine learning, 45* (2001), 5-32.
3. Borg, I. and Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.
4. Cristianini, N. and Shawe-Taylor, J. *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press, 1 edition, March 2000.
5. Hen, I., Sakov, A., Kafkafi N., Golani, I., Benjamini, Y. The dynamics of spatial behavior: how can robust smoothing techniques help? Journal *of Neuroscience Methods, 133* (2004), 161-172.
6. Hodges, A., Hughes G., Brooks, S., Elliston, L., Holmans, P., Dunnett, S. and Jones, L. Brain gene expression correlates with changes in behavior in the r6/1 mouse model of Huntington's disease. *Genes, Brain and Behavior 7* (2008), 288-299.
7. Jollie, I.T. *Principal Component Analysis*. Springer, second edition, October 2002.
8. Martiskainen, P., Jarvinen, M., Skon, J., Tiirikainen, J., Kolehmainen, M. and Mononen J. Cow behaviour pattern recognition using a three- dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science, 119* (2009): 32-38.
9. Noldus, L., Grieco, F.,Loijens, L., and Zimmerman, P. (Eds.). *Measuring behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 2005*.

10. Rakotomamonjy, A. Variable selection using svm based criteria. *The Journal of Machine Learning Research*, *3* (2003), 1357–1370.
11. Scholkopf, B. and Smola, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.
12. Spruijt, B.M. and De Visser, L., Advanced behavioural screening: automated home cage ethology. *Drug Discovery Today: Technologies*, *3* (2006), 231-237.
13. Van Vliet, S., Vanwersch R., Jongsma M., Van der Gugten, J., Olivier, B. and Philippens, I. Neuroprotective effects of modanil in a marmoset Parkinson model: behavioral and neurochemical aspects. Behavioural Pharmacology, 17(5-6):453{462, 2006. Professional ethics. Social Science Computing Review *10*, 2 (1992), 453-469.
14. Vapnik ,V. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
15. Zou, H. and Hastie, T. , Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67* (2005), 301.