Ladds *et al. Anim Biotelemetry (2017) 5:8*
DOI 10.1186/s40317-017-0123-1

Animal Biotelemetry

TELEMETRY CASE REPORT

Open Access

Super machine learning: improving accuracy and reducing variance of behaviour classification from accelerometry

Monique A. Ladds[1*], Adam P. Thompson[2], Julianna-Piroska Kadar[3], David J Slip[1,4], David P Hocking[5] and Robert G Harcourt[1]

Abstract

Background: Semi-automating the analyses of accelerometry data makes it possible to synthesize large data sets. However, when constructing activity budgets from accelerometry data, there are many methods to extract, analyse and report data and results. For instance, machine learning is a robust approach to classifying data. We used a new method, super learning, that combines base learners (different machine learning methods) in an optimal manner to achieve overall improved accuracy. Other facets of super learning include the number of behavioural categories to predict, the number of epochs (sample window size) used to split data for training and testing and the parameters on which to train the models.

Results: The super learner accurately classified behaviour categories with higher accuracy and lower variance than comparative models. For all models tested, using four behaviours, in comparison with six, achieved higher rates of accuracy. The number of epochs chosen also affected the accuracy with smaller epochs (7 and 13) performing better than longer epochs (25 and 75).

Conclusions: Correct model selection, training and testing are imperative to creating reliable and valid classification models. To do so means model fitting must use a wide array of selection criteria. We evaluated a number of these including model, number of behaviours to classify and epoch length and then used a parameter grid search to implement the models. We found that all criteria tested contributed to the models' overall accuracies. Fewer behaviour categories and shorter epoch length improved the performance of all models tested. The super learner classified behaviours with higher accuracy and lower variance than other models tested. However, when using this model, users need to consider the additional human and computational time required for implementation. Machine learning is a powerful method for classifying the behaviour of animals from accelerometers. Care and consideration of the modelling parameters evaluated in this study are essential when using this type of statistical analysis.

Keywords: Behavioural classification, Marine mammal, Ethogram, Accelerometer, Machine learning, Super learner

Background

Advances in biologging technologies and computing power have enabled biologists to pry into the daily existence of many difficulties to observe animals [1, 2]. A powerful new approach is to create ethograms from accelerometers using machine learning [3]. Accelerometers measure the inertial acceleration of an animal while moving, most commonly on three axes [4]. Unique combinations of these three axes over a period of time identify specific movements that correspond to a single behaviour or series of behaviours. Binary classes of behaviour can be identified with high degree of accuracy using machine learning, e.g. prey captures in penguins using support vector machines (SVM's) [5]. A variety of machine learning algorithms have attempted to distinguish between multiple classes of behaviours, some

*Correspondence: monique.ladds@hdr.mq.edu.au
[1] Marine Predator Research Group, Department of Biological Sciences, Macquarie University, North Ryde, NSW 2113, Australia
Full list of author information is available at the end of the article

successfully [e.g. 6–10] and some with less success [e.g. 11]. There are a number of reasons why machine learning methods may not be able to classify data from accelerometry accurately, including the number of categories to be classified [12], the duration of the sample of behaviour to be classified [13], the number of sample behaviours to classify from [14] and the machine learning method that is used [7]. Using too many categories of behaviour may affect the ability of the machine learning method to accurately classify all behaviour. For example, the attack/peck category created for crab plovers could not be predicted using decision trees from a study attempting to classify seven categories of behaviour [11]. Using hidden semi-Markov models, two categories of behaviours were able to be classified with much higher accuracy than three, four or five categories [12]. Accuracies of machine learning models are likely to improve when using fewer categories because the algorithm has fewer classes to distinguish between. This is especially true if the classes that are being combined are often misclassified as each other. For example, if we have three classes A, B and C, and classes B and C are often misclassified as each other, then combining them into one class will increase classification accuracy, at the cost of less detail overall. Alternatively, extending the sample of time from the accelerometry data used to classify behaviour can improve the overall accuracy machine learning methods by providing more samples overall [13].

The machine learning method selected to classify the data will also influence the overall accuracy [7, 12]. There have been several attempts to evaluate the accuracies of different machine learning methods [7, 13, 15]. However, due to vastly distinct dynamic movement of different animal species, it is unlikely that there will ever be a universal set template for creating ethograms from accelerometry [16, 17]. Instead, a new machine learning method described here may afford a solution to the problem of method selection. Super learning takes a set of candidate learners (other machine learning methods), applies them to a data set and chooses an optimal learner or combination of learners based on the resultant cross-validated risk [18]. The super learner model (SL) seeks to find the optimal combination candidate learners such that it will perform as well or better than any of the learner inputs [19]. Super learning has previously been applied to large medical data sets in order to make survival predictions with considerable success [20], but has until now not been evaluated for its ability to classify behaviour from accelerometry data.

The ability to reliably build highly generalizable models for the classification of animal behaviour will be a significant advance for the study of those species that are difficult or impossible to observe in the wild or sustain in captivity [6, 16]. Otariid pinnipeds, fur seals and sea lions, play an important role in the trophic interactions of many marine ecosystems [21], yet despite the importance of this group to understanding marine ecosystems, there is still much to learn about the behaviour of these and other marine predators [22]. Marine animals are very difficult to observe in the wild as they are active in remote locations and deep underwater where direct observation is often not possible, but being large and semi-aquatic, otariids are ideal candidates for remote observation using accelerometry [2, 23]. To reliably classify animal behaviours from accelerometry, it is necessary to evaluate the performance of different models and their parameters [7]. The aims of this study are twofold: (1) assess whether super learning can improve the accuracy of classifying accelerometry data in general and (2) identify the optimal time window and number of behaviour categories required to create reliable ethograms for a representative group of animals: fur seals and sea lions.

Methods

Animals

We conducted captive experiments at three Australian marine facilities: Dolphin Marine Magic, Coffs Harbour (RF1: −30°17′N, 153°8′E); Underwater World, Sunshine Coast (RF2: −25°40′N, 153°7′E); and Taronga Zoo, Sydney (RF3: −33°50′N, 151°14′E) from August to November 2014 and again at RF2 in August 2015. We used two Australian fur seals (*Arctocephalus pusillus doriferus*), three New Zealand fur seals (*Arctocephalus forsteri*), one subantarctic fur seal (*Arctocephalus tropicalis*) and six Australian sea lions (*Neophoca cinerea*) (Table 1). All seals were on permanent display at their respective marine facilities and were fed and cared for under the guidelines of the individual facility. All Australian sea lions in the study were born as part of an ongoing captive breeding programme in Australian aquaria. All fur seals came into captivity as juveniles after they were found in poor health or were injured and being deemed unsuitable for release back into the wild.

Experimental protocol

We used a triaxial accelerometer (CEFAS G6a+: 40 mm × 28 mm × 16.3 mm, 18 g in air and 4.3 g in seawater, CEFAS Technology Ltd, Lowestoft, UK) to measure the movement of the seals. We used two attachment methods for accelerometers: either taped between the shoulder blades or secured in a custom-designed harness. Accelerometers were set to record at ±8 g and at 25 samples per second (25 Hz) on each axis. We recorded all trials continuously with one or two cameras (GoPro Hero 3—Black edition, USA; HDRSR11E: Sony, Japan), and trials had a maximum duration of 2.5 h. Videos were scored

Table 1 Study species and characteristics of seal identification, marine facility, species, age, mass range, sex, number of trials and method of accelerometer attachment for fur seals and sea lions used in the study

Seal ID	Marine facility	Species	Age	Mass range (kg)	Sex	Number of trials	Attachment method
ASF1	RF1	ASL	5	44–47	Female	13	Harness
ASF3	RF2	ASL	17	58–74	Female	4	Harness
ASF4	RF1	ASL	17	66–70	Female	12	Harness
ASF6	RF1	ASL	7	50	Female	2	Harness
ASM1	RF1	ASL	9	108–110	Male	8	Harness
AFF1	RF2	AFS	17	69–79	Female	7	Tape
AFM1	RF2	AFS	16	175–242	Male	7	Tape
ASM2	RF3	ASL	13	160–162	Male	9	Tape
NFM1	RF3	NZFS	8	47–54	Male	5	Tape
NFM2	RF2	NZFS	11	108–152	Male	5	Tape
NFM3	RF3	NZFS	13	111–154	Male	8	Tape
SFM1	RF2	SFS	4	28–30	Male	3	Tape

AFS Australian fur seal, *NZFS* New Zealand fur seal; *SFS* subantarctic fur seal, *ASL* Australian sea lion

to an ethogram consisting of 26 unique behaviours developed previously [14]. We time-matched the videos and the accelerometry output to generate annotated acceleration data sets.

Behaviour segmenting

We grouped the 26 behaviours into broader behavioural categories. As the number of behavioural categories used to classify behaviour may affect the overall results, the analysis was run twice using four (feeding, grooming, resting and travelling) and then six categories (feeding, foraging, thrashing, grooming, resting and travelling) (Table 3; for a description of the individual behaviours in each of the categories please see [14—S1 File]). We also compared the ability of the model to discriminate behaviours over a range of discrete periods. We tested four epochs (number of accelerometer samples): 7 (0.28 s), 13 (0.52 s), 25 (1 s) and 75 (3 s) [24]. Behaviours could also be "contaminated" where two behaviours occur in the same time window. In these cases, we used the dominant behaviour with resultant windows of uneven time duration.

Summary statistics

We created 147 summary statistics as the inputs to the machine learning models. Most were summary statistics created from the *x*, *y* and *z* inputs (described below), and a few related to the animal or the behaviour including where the behaviour occurred (surface, underwater or land), device attachment method (harness or tape), age, mass, sex and species of the individual [14]. The location of the behaviour was determined by observation; however, in the wild, it can be using a combination of depth and the wet/dry sensor on the accelerometer (M. Ladds, M. Salton, R. McIntosh, D. Hocking, D. Slip, R. Harcourt, unpublished observations). For each of the three axes (*x*, *y*, *z*), we calculated mean, median, minimum, maximum, range, standard deviation, skewness, kurtosis, absolute value, inverse covariance and autocorrelation trend (the coefficient derived from a linear regression) and the 10th and 90th percentiles. We also calculated *q* as the square root of the sum of squares of the three axis [7] and included pairwise correlations of the three axis (*x*–*y*, *y*–*z*, *x*–*z*) [25]. The inclination and azimuth were calculated as per Nathan et al. [7]. We calculated dynamic body acceleration (DBA) by using a running mean of each axis over three seconds to create a value for static acceleration [26]. We then subtracted the static acceleration at each point from the raw acceleration value to create a value for partial dynamic body acceleration (PDBA). We calculated overall dynamic body acceleration (ODBA) [26, 27] using

$$\mathrm{ODBA} = \left|X_{\mathrm{dyn}}\right| + \left|Y_{\mathrm{dyn}}\right| + \left|Z_{\mathrm{dyn}}\right| \quad (1)$$

We calculated vectorial dynamic body acceleration (VeDBA) [28] using

$$VeDBA = \sqrt{X_{dyn}^2 + Y_{dyn}^2 + Z_{dyn}^2} \quad (2)$$

We calculated the area under the curve for both ODBA and VeDBA using the package "MESS" in R [29, 30]. The minimum, maximum and 10th and 90th percentiles were calculated for PDBA, ODBA and VeDBA.

Classification models

There are many candidate models suitable for classifying behavioural data obtained from accelerometry [7], and choosing the most appropriate method for the data in question can be complicated and time-consuming. The

super learner model (SL) combines candidate models (other machine learning models, henceforth referred to as base learners) by applying a selection of them to a set of data and then weighting all of these learners through another learner. The optimal combination is chosen based on cross-validated risk [18, 31]. The base learners chosen for this study were: random forests (RF), gradient boosting machine (GBM) (both of which have previously been demonstrated to effectively classify this type of data well [14]) and a baseline model, logistic regression (LR) to which performances of the other models could be compared. Logistic regression was included as a baseline model as it is well tested, easy to implement and unlikely to overfit. Each base learner was trained across a set of parameters, with the predictions of each model kept. These predictions, plus the raw data, then became the inputs to the SL. The SL then learned from the predictions of the base learners as well as the summary and feature statistics to predict the outcomes.

For each of the models, data were split into a train (evaluation) and test (validation) set using 70 and 30% of the data, respectively. In total, the models were trained on ~90,000 individual data points or roughly ~13 h of coded data. Note that the test data were not seen by the model during training. This ensured that the scores obtained from the models reflected the ability of the model to predict from data outside training. Results of the model were reported as cross-validation scores and out-of-sample scores, which include accuracy and kappa (Additional file 1). Accuracy was the proportion of true positives identified by the model, while kappa was employed as more than two observers were used to classify data, thereby providing a measure for the fact that some of their observations will agree or disagree by chance [32]. This value was used to assess agreement of observed and predicted values in the confusion tables [24]. Precision and sensitivity are reported in the confusion matrix (Table 4) where precision is defined as the proportion of predictions from a behaviour category that were actually that behaviour, and sensitivity is the proportion of behaviours from a category that were classified as that behaviour [16].

Parameter grid search

Within each model, there were a number of parameters from which models can be trained. Samples of each of these parameters were chosen, and each model was run through every combination using a grid search (Table 2; Additional file 2). We evaluated best parameter grids of each model using H20 [33] for GBM and RF, glmnet [34] for LR and the SL. All analyses were run using R [30].

Results

Triaxial acceleration data were collected from 12 seals over a range of trials lasting in duration from 10 min to 2.5 h (Table 1). From these we were able to mark 7525 bouts of behaviour, split into either four or six categories (Table 3).

Comparing model performance

All three test models (SL, RF and GBM) had significantly higher accuracies across the range of epochs and categories of behaviour tested compared to the baseline model (LR; Fig. 1). The SL accuracy ranged from 71.6% (7 epochs) to 73.6% (13 epochs) accuracy for six categories of behaviour and from 83.4% (25 and 75 epochs) to 85.1% (13 epochs) accuracy for four categories of behaviour (Additional file 2). The RF achieved slightly less accuracy ranging from 82.3% (75 epochs) to 84.4% (13 epochs) for four categories and from 67.8% (75 epochs) to 72.7% (13 epochs) for six categories. GBM performed slightly less well than the SL and about the same as the RF with accuracies ranging from 70.9% (75 epochs) to 73.4% (13 epochs) for six behaviour categories and from 82.0% (75 epochs) to 84.7% (13 epochs) for four categories of behaviour. The LR accuracies were significantly below all

Table 2 Parameters for the four models tested

Model	Nbins	Mtry	Ntree	Max depth
Random forest	20, 30, 40 (numeric) 3 (categorical)	5, 10, 15	200	5, 10, 15

	Nbins	Learn rate	ntree	Max depth	Sample rate
Gradient boosting machine	20, 30, 40 (numeric) 3 (categorical)	0.1, 0.001	250, 700	5, 10	0.7, 0.8, 0.9

	Lambda	Alpha
Logistic regression and super learner	Range exp(−11) to exp(6)	0–1 by 0.025

Nbins number of bins, *Mtry* number of splits in branches, *Ntree* total number of trees grown, *Max depth* maximum depth to grow the trees (for a detailed description of the model parameters and how they are used see Additional file 3)

Table 3 Number of unique behaviours observed from video analysis for each category of behaviour

Four categories	Six categories	Behaviour	Number of bouts
[1]		Walking	545
Travelling		Surface swimming	1133
(*N* = 2864)		Swimming	1008
		Fast	121
		Porpoising	57
[2]		Lying	17
Resting		Sitting	541
(*N* = 839)		Still	281
[3]	[3]	Scratch	68
Grooming	Grooming	Rubbing	10
(*N* = 334)	(*N* = 245)	Sailing	29
		Jugging	19
		Face rub	54
		Rolling	115
	[NA] High frequency	Shake	39
[4]	[4]	Chewing	309
Feeding	Feeding	Manipulation	792
(*N* = 1841)	(*N* = 1615)	Capture	394
		Hold and tear	120
	[5] Foraging (*N* = 226)	Searching	226
[6] Thrashing (*N* = 303)		Thrashing	303
Other	Playing		30
(*N* = 1344)	In/out		475
	Other		839

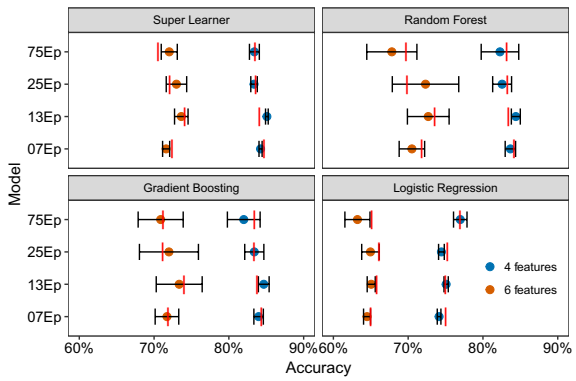

Fig. 1 Classification accuracy from cross- and out-of-sample validation of four different machine learning algorithms. Coloured points (*blue* four-feature models; *orange* six-feature models) represent out-of-sample accuracy with*error bars* of ±1 SD. *Red bars* represent cross-validation accuracy for each associated model. *SL* super learner, *RF* random forest, *GBM* gradient boosting model, *LR* logistic regression

of these for all categories ranging from 74.1% (7 epochs) to 77.0% (75 epochs) for four categories and from 63.2% (75 epochs) to 65.1% (13 epochs) for six categories.

SL classified categories of behaviour with higher accuracy and lower variance than both RF and GBM across all epochs (except GBM 7 epochs, six categories). The variance was reduced by ~70% across all model combinations tested, and accuracy was improved by between −0.1 and 10.1% (Fig. 1; Additional file 1). The variances obtained from the logistic regression models were similar to the SL. Accuracy and precision of all models improved when using four as opposed to six categories of behaviour. Looking at the overall performance of the models from the highest cross-validation score, out-of-sample score and the kappa score, we concluded that using 13 epochs produced the best results across the four models (Additional file 1).

Identifying categories of behaviour

Across all models and epochs, grooming and resting classified with the highest accuracy, with grooming generally outperforming resting (Fig. 2; Additional file 2). Examining the confusion matrix from the best performing model (SL—four behaviours, 13 epochs), the classification errors from the four categories of behaviour revealed that foraging often misclassified as travelling and vice versa (Table 4). Overall, within the test models (SL, RF, GBM), all four behaviours were correctly classified more than 75% of the time (Fig. 2). Within the six behaviour categories, the main misclassification stemmed from feeding, where only the super learner classified it correctly more than 50% of the time. The "thrashing" category that was also added to the model was classified with high accuracy (>75%). Resting and grooming maintained their high predictive accuracies across the test models (>80%). Foraging also maintained a reasonably high rate of classification (>70%), while travelling lost around 10% accuracy when compared with the four behaviour models.

Discussion

The aim of this study was to assess whether super learning would improve the predictive ability of base learners (RF, GBM and LR) to classify behaviour from free-living animals using accelerometry. While building machine learning models, a number of choices must be considered about how to segment the data. We evaluated several combinations of time segmentation and number of behaviour categories for this type of accelerometry data. Using super learning increased the accuracy of the models, albeit only slightly, and reduced the prediction error when compared with RF, GBM and the baseline model—LR. Shorter time windows (<13 samples) and fewer categories of behaviour (4 vs. 6) were better at predicting the behaviour state.

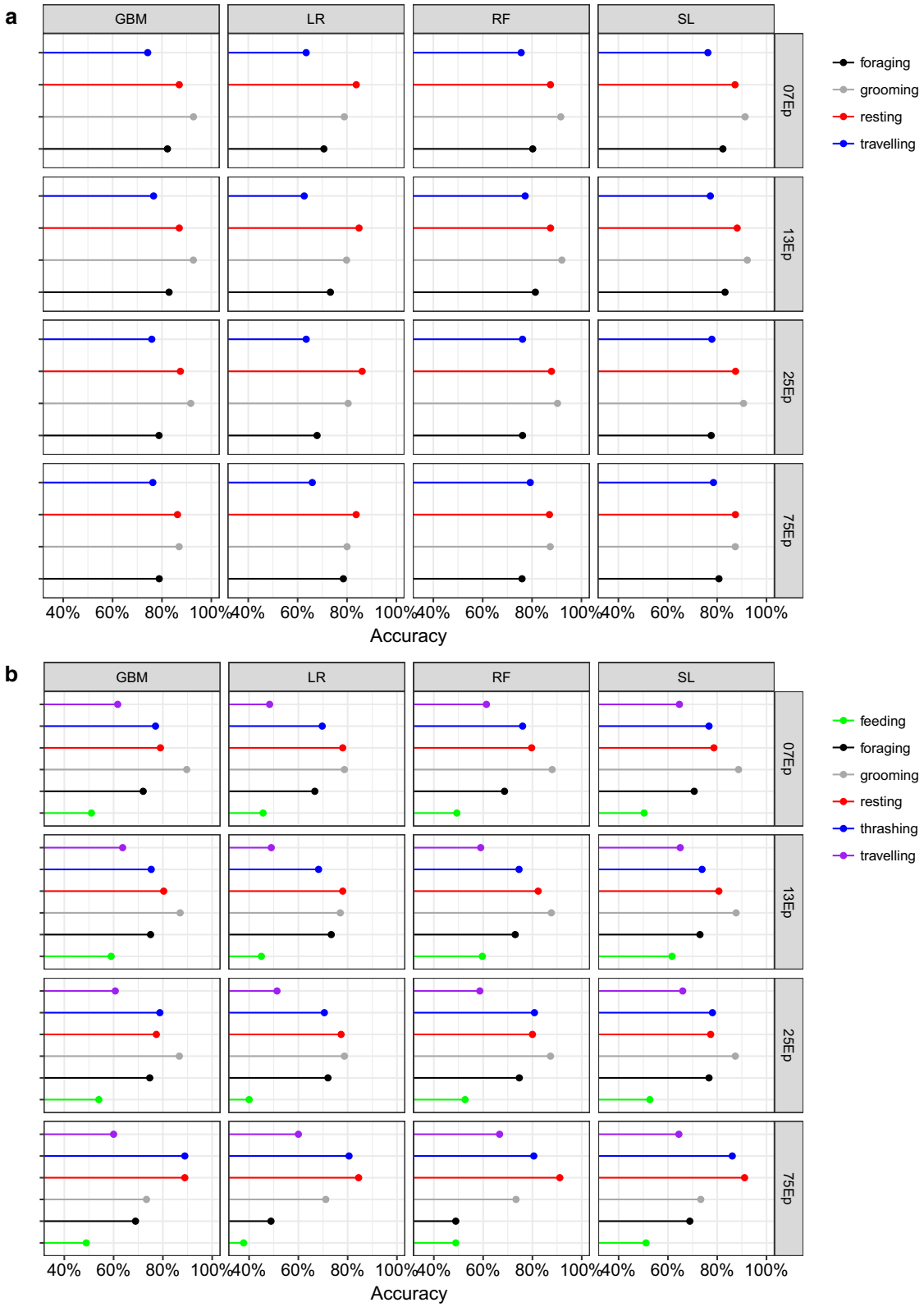

Fig. 2 Classification accuracy of behaviour across epochs and models. Four (**a**) and six (**b**) categories of behaviour were tested across four (SL, RF, GBM and LR) models across four (7, 13, 25, 75) epochs

Table 4 Confusion matrices from three test models using four behaviours and 13 epochs

	Foraging	Grooming	Resting	Travelling	Precision	Sensitivity
Super learner						
Foraging	1248	17	53	182	0.83	0.82
Grooming	18	1292	27	64	0.92	0.91
Resting	80	37	1321	61	0.88	0.89
Travelling	185	79	77	1158	0.77	0.79
Gradient boosting machine						
Foraging	1243	23	54	180	0.83	0.81
Grooming	20	1300	30	52	0.93	0.89
Resting	80	39	1305	76	0.87	0.90
Travelling	191	92	68	1149	0.77	0.79
Random forest						
Foraging	1220	25	57	198	0.81	0.80
Grooming	17	1291	42	52	0.92	0.90
Resting	86	35	1312	67	0.87	0.89
Travelling	195	88	59	1158	0.77	0.79

Number of behaviour categories: Less is more?

Four behavioural categories had a higher classification rate than six behaviours. At its most basic, accelerometers discriminate between two behavioural states (e.g. activity vs. resting or swimming vs. prey capture) and can do so accurately [5, 35]. Adding more categories for the model to discriminate increases complexity, but reduces the uniqueness of the model, thus decreasing its overall accuracy [12, 13]. There is also a greater chance of overlap with other behavioural categories. Increasing behaviour categories from four to six produced an overall average 11.5% (range 9.5–14.5%) decrease in accuracy. The optimal number of categories becomes a trade-off between useful ecological information and high accuracy. Reducing the number of categories broadens the scope of the remaining categories as more similar behaviours are considered together and are thus easier to discriminate by the model. An important distinction to make is that considering fewer categories does not mean removing behaviours from the models, because if those behaviours are observed in the wild, the model will still try to classify them, resulting in an inaccurate representation of what the animal did while being monitored (for a discussion of this issue see [14]). As the loss in accuracy is so small, this leaves it up to the researcher to determine whether quality (fewer behaviours—more accuracy) or quantity (more behaviours—less accuracy) is important in the study. In this illustration of the method, which is broadly applicable to all free-living animals that can be equipped with accelerometers, we used fur seals and sea lions. For species such as these, four behavioural categories appear to be the minimum that provides meaningful information about their activities. In future studies that use this method, the number of categories must be tailored to the species concerned and aims of the study.

Epoch size: Smaller is better?

We found that smaller epochs gave better overall predictions, and that the length of the epoch was significant in predicting different categories of behaviour. Increasing the window size reduces the sample size, which likely decreases the overall ability of the models to predict accurately. Having smaller epochs increases the sample size and reduces the chances of the model overfitting. Large epoch sizes are also more likely to capture more than one behaviour, increasing the difficulty for the model to distinguish between classes. In contrast to our results, a study of cow behaviour found that longer epochs tended to perform better than shorter epochs (5 and 10 vs. 1 min) [13]. However, a similar study with humans discovered that epochs of one to two seconds had the best precision values [36]. They also found that epoch length significantly affected the overall accuracy of individual behaviours, which concords with our findings. We found different prediction accuracies by adding thrashing and feeding to the model. All models predicted thrashing with high accuracy (~75%), while only the SL predicted feeding with more than 50% accuracy. Thrashing is a very distinctive behaviour, with accelerometer readings exceeding 4 g; very few other behaviours have this feature. By contrast, we defined feeding as a seal taking fish out of the water column, and animals were swimming while taking fish; therefore, it was difficult for the models to distinguish between these two behaviours. Therefore, any additional behaviours added to the base four-category model need to be clearly distinct from any

other behaviour. Future studies investigating seal feeding behaviour should seek to gather examples of seals capturing live prey.

Super models: Is it worth it?

The idea of a super machine learning model is enticing, allowing a multitude of machine learning models to be trained and tested on a single set of data and thus allowing the model to optimally combine each of the individual models to give better overall predictions. Super learning has been successfully used in medical research [20] and spatial analyses [19] and improved the behaviour classification models from accelerometry, albeit marginally. With the exception of a single model combination (GBM; 7 epochs, 6 features), the super learner performed better than any other model combination. This was expected as super learning will use the optimal model it has trained on if it is unable to compute a more optimal solution [19]. We found an average increase of 3.4% (range −0.1 to 10.1%) in the classification accuracy of the models using super learning. While any improvement in model performance is welcome, single-state-of-the-art algorithms like GBM are easy to implement in software environments like R. However, this research has only investigated a small aspect of the potential of super learner models. Super learners are unrestricted by the number and type of models that constitute the base learners, so can be optimized for the type of data that is input. This is particularly useful if researchers are interested in a particular behaviour that is usually difficult to distinguish with a single model (i.e. attack/peck in plovers [11]) or very high accuracy is imperative for the research objectives. We suggest the individual researcher takes this into consideration when deciding whether the additional human and computational time required to implement super learning will be beneficial for their behavioural data study.

Conclusions

This study evaluated a number of machine learning methods to classify accelerometry data and compare them to a new method—super learning. We found that super learning improved the accuracy and reduced the variance in the predictive accuracy of the model. We showed that the epochs (number of samples) and number of behavioural categories influenced the overall accuracy of the model. This study demonstrates the importance of evaluating all options when using machine learning to classify animal behaviour. While this is by no means an exhaustive demonstration of the possible choices to be made when implementing machine learning methods, the options highlighted here (number of behaviour categories, epoch size, model selection and parameter grid search) are some of the most important and easiest to test when conducting this type of statistical analysis. Future studies classifying animal behaviour from accelerometry using machine learning should, where possible, test their models across a selection of these options in order to obtain the highest accuracies.

Additional files

Additional file 1. Accuracy summaries and sample size used to train and test for four machine learning models. Models were tested across four different size epochs and with four and six behavioural categories. Statistics reported are: cross validation (training) accuracy and 95% confidence interval; out-of-sample (testing) accuracy and standard deviation (SD); the Kappa statistic and the proportion improvement made by the SL compared to other models.

Additional file 2. Sensitivity and specificity of each behaviour category for all model combinations tested. *SL* super learner, *RF* random forest, *GBM* stochastic gradient boosting, *LR* logistic regression.

Additional file 3. Parameter grid search variables.

Abbreviations

RF: random forest; GBM: gradient boosting machine; SL: super learner; LR: logistic regression; PDBA: partial dynamic body acceleration; ODBA: overall dynamic body acceleration; VeDBA: vectorial dynamic body acceleration.

Authors' contributions

ML conceived the study design, collected data, performed data analysis and drafted the manuscript. AT performed data analysis and contributed to writing. DH, RH and DS were involved in study design, data collection and manuscript editing. JK was involved in data collection, data analysis and manuscript editing. All authors read and approved the final manuscript.

Author details

[1] Marine Predator Research Group, Department of Biological Sciences, Macquarie University, North Ryde, NSW 2113, Australia. [2] TAL Life Limited, La Trobe St, Melbourne, VIC 3000, Australia. [3] Taronga Conservation Society Australia, Bradley's Head Road, Mosman, NSW 2088, Australia. [4] Ecology of Fishes Lab, Department of Biological Sciences, Macquarie University, North Ryde, NSW 2113, Australia. [5] School of Biological Sciences, Monash University, Melbourne, VIC, Australia.

Acknowledgements

We thank all of the marine mammal staff at Dolphin Marine Magic, Underwater World Mooloolaba and Taronga for their invaluable assistance with data collection, training the seals and ongoing commitment to this project. We thank Guy Bedford for his assistance in designing and producing the harness used for the sea lions.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data set(s) supporting the conclusions of this article is(are) available in the GitHub repository, https://github.com/aptperson/behaviour_accelerometry.

Ethics approval

This study was carried out under permits from Macquarie University Ethics Committee (ARA-2012_064) and Taronga Ethics Committee (4c/10/13). All experiments were conducted under the current laws of Australia authorized under New South Wales Office of Environment and Heritage Scientific Licence SL100746 to RH.

Funding

This project is funded by Australian Research Council Linkage Grant (Grant Number LP110200603) to RH and DS, with support from Taronga Conservation Society, Australia. ML is a recipient of a Macquarie University Research

Excellence Scholarship. Funding was provided by Macquarie University (Grant No. 43000215).

Received: 27 September 2016 Accepted: 9 March 2017
Published online: 29 March 2017

References

1. Wilson RP, Shepard E, Liebsch N. Prying into the intimate details of animal lives: use of a daily diary on animals. Endanger Species Res. 2008;4:123–37. doi:10.3354/esr00064.
2. Hussey NE, Kessel ST, Aarestrup K, Cooke SJ, Cowley PD, Fisk AT, Harcourt RG, Holland KN, Iverson SJ, Kocik JF, et al. Aquatic animal telemetry: a panoramic window into the underwater world. Science. 2015;348:1255642. doi:10.1126/science.1255642.
3. Sakamoto KQ, Sato K, Ishizuka M, Watanuki Y, Takahashi A, Daunt F, Wanless S. Can ethograms be automatically generated using body acceleration data from free-ranging birds? PLoS ONE. 2009;4:e5379. doi:10.1371/journal.pone.0005379.
4. Brown DD, Kays R, Wikelski M, Wilson R, Klimley AP. Observing the unwatchable through acceleration logging of animal behavior. Anim Biotelem. 2013;1:20. doi:10.1186/2050-3385-1-20.
5. Carroll G, Slip DJ, Jonsen I, Harcourt RG. Supervised accelerometry analysis can identify prey capture by penguins at sea. J Exp Biol. 2014;217:4295–302. doi:10.1242/jeb.113076.
6. Bidder OR, Campbell HA, Gomez-Laich A, Urge P, Walker J, Cai YZ, Gao LL, Quintana F, Wilson RP. Love thy neighbour: automatic animal behavioural classification of acceleration data using the K-nearest neighbour algorithm. PLoS ONE. 2014;9:7. doi:10.1371/journal.pone.0088609.
7. Nathan R, Spiegel O, Fortmann-Roe S, Harel R, Wikelski M, Getz WM. Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. J Exp Biol. 2012;215:986–96. doi:10.1242/jeb.058602.
8. Resheff YS, Rotics S, Nathan R, Weinshall D. Matrix factorization approach to behavioral mode analysis from acceleration data. In: 2015 IEEE international conference on data science and advanced analytics (DSAA), 19–21 October 2015; 2015. p. 1–6.
9. Resheff YS, Rotics S, Harel R, Spiegel O, Nathan R. AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements. Mov Ecol. 2014;2:27. doi:10.1186/s40462-014-0027-0.
10. Chimienti M, Cornulier T, Owen E, Bolton M, Davies IM, Travis JMJ, Scott BE. The use of an unsupervised learning approach for characterizing latent behaviors in accelerometer data. Ecol Evol. 2016;6:727–41. doi:10.1002/ece3.1914.
11. Bom RA, Bouten W, Piersma T, Oosterbeek K, van Gils JA. Optimizing acceleration-based ethograms: the use of variable-time versus fixed-time segmentation. Mov Ecol. 2014;2:1–8. doi:10.1186/2051-3933-2-6.
12. Hammond TT, Springthorpe D, Walsh RE, Berg-Kirkpatrick T. Using accelerometers to remotely and automatically characterize behavior in small animals. J Exp Biol. 2016;219:1618–24. doi:10.1242/jeb.136135.
13. Diosdado JAV, Barker ZE, Hodges HR, Amory JR, Croft DP, Bell NJ, Codling EA. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. Anim Biotelem. 2015;. doi:10.1186/s40317-015-0045-8.
14. Ladds MA, Thompson AP, Slip DJ, Hocking DP, Harcourt RG. Seeing it all: evaluating supervised machine learning methods for the classification of diverse otariid behaviours. PLoS ONE. 2016;11:e0166898. doi:10.1371/journal.pone.0166898.
15. Dutta R, Smith D, Rawnsley R, Bishop-Hurley G, Hills J, Timms G, Henry D. Dynamic cattle behavioural classification using supervised ensemble classifiers. Comput Electron Agric. 2015;111:18–28. doi:10.1016/j.compag.2014.12.002.
16. Campbell HA, Gao L, Bidder OR, Hunter J, Franklin CE. Creating a behavioural classification module for acceleration data: using a captive surrogate for difficult to observe species. J Exp Biol. 2013;216:4501–6. doi:10.1242/jeb.089805.
17. Gerencser L, Vasarhelyi G, Nagy M, Vicsek T, Miklosi A. Identification of behaviour in freely moving dogs (*Canis familiaris*) using inertial sensors. PLoS ONE. 2013;8:e77814. doi:10.1371/journal.pone.0077814.
18. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007;6. Article 25.
19. Davies MM, van der Laan MJ. Optimal spatial prediction using ensemble machine learning. Int J Biostat. 2016;12:179–201.
20. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3:42–52. doi:10.1016/S2213-2600(14)70239-5.
21. Bowen W. Role of marine mammals in aquatic ecosystems. Mar Ecol Prog Ser. 1997;158:267–74.
22. Hays GC, Ferreira LC, Sequeira AMM, Meekan MG, Duarte CM, Bailey H, Bailleul F, Bowen WD, Caley MJ, Costa DP, et al. Key questions in marine megafauna movement ecology. Trends Ecol Evol. 2016;31:463–75. doi:10.1016/j.tree.2016.02.015.
23. Bowen W, Tully D, Boness D, Bulheier B, Marshall G. Prey-dependent foraging tactics and prey profitability in a marine mammal. Mar Ecol Prog Ser. 2002;244:235–45.
24. Alvarenga FAP, Borges I, Palkovič L, Rodina J, Oddy VH, Dobos RC. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. Appl Anim Behav Sci. 2016;181:91–9. doi:10.1016/j.applanim.2016.05.026.
25. Ravi N, Dandekar N, Mysore P, Littman ML. Activity recognition from accelerometer data. In: Proceedings of the seventeenth conference on innovative applications of artificial intelligence, July 9–13; Pittsburgh; 2005. p. 1541–1546.
26. Shepard EL, Wilson RP, Halsey LG, Quintana F, Laich AG, Gleiss AC, Liebsch N, Myers AE, Norman B. Derivation of body motion via appropriate smoothing of acceleration data. Aquat Biol. 2008;4:235–41. doi:10.3354/ab00104.
27. Wilson RP, White CR, Quintana F, Halsey LG, Liebsch N, Martin GR, Butler PJ. Moving towards acceleration for estimates of activity specific metabolic rate in free living animals: the case of the cormorant. J Anim Ecol. 2006;75:1081–90. doi:10.1111/j.1365-2656.2006.01127.x.
28. Qasem L, Cardew A, Wilson A, Griffiths I, Halsey LG, Shepard EL, Gleiss AC, Wilson R. Tri-axial dynamic acceleration as a proxy for animal energy expenditure; should we be summing values or calculating the vector? PLoS ONE. 2012;7:e31187. doi:10.1371/journal.pone.0031187.
29. Ekstrom C. MESS: miscellaneous esoteric statistical scripts. In: R package version 03-2 R package version 0.3-2 edition; 2014.
30. R Core Development Team. R: a language and environment for statistical computing. In: R version 331, R package version 3.2.3 edition. Vienna: R Foundation for Statistical Computing; 2015.
31. Sinisi SE, Polley EC, Petersen ML, Rhee S-Y, van der Laan MJ. Super Learning: an application to the prediction of HIV-1 drug resistance. Stat Appl Genet Mol Biol 2007;6. Article 7. doi:10.2202/1544-6115.1240.
32. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med. 2005;37:360–3.
33. Lendell E. h2oEnsemble. 2015.
34. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
35. Takahashi M, Tobey JR, Pisacane CB, Andrus CH. Evaluating the utility of an accelerometer and urinary hormone analysis as indicators of estrus in a zoo-housed koala (*Phascolarctos cinereus*). Zoo Biol. 2009;28:59–68. doi:10.1002/zoo.20212.
36. Huynh T, Schiele B. Analyzing features for activity recognition. In: Proceedings of the 2005 joint conference on smart objects and ambient intelligence. New York: Association for Computing Machinery; 2005. p. 159–163.