University of Amsterdam

Leren en beslissen project

Examining the Oysterchatcher's timebudgets using clustering on accelerometer data

Authors:
Sebastian Dröpelmann
Didier Gumbs
Frank Smit
Sander Latour

Supervisor:
Dr. Maarten van Someren
Prof. Dr.ir. Willem Bouten

February 4, 2011

Contents

1 Introduction

In modern biology accelerometers are being used to get further understanding of animal behavior and how an animal divides its time. Accelerometers are useful because they are fairly inexpensive and do not interfer with the animals behavior.

In this paper we have turned to the Oystercatcher *Haematopus ostralegus*, a shorebird which can be found on coasts worldwide. The data we received was taken from 57 birds which live on Schiermonnikoog, an island of The Netherlands, this data will be discussed in further sections. With this data we will try to find certain sets which match and form a cluster. This will be achieved with machine learning techniques and adding new features which are derived from the accelerometer data. We have focused on unsupervised classification due to the data of the accelerometer which is unlabeled and manually labeling the data from the accelerometer is exhaustive because of the size of the data.

Willem Bouten a researcher at the Institute for Biodeversity an Ecosystem Dynamics, who leads a research group of Computational Geo-Ecology, would like to combine the knowledge of his research with artificial intelligence. As our contact for our research he helped us understand different means of interpreting the data aswell as Oystercatcher behavior. He also provided us with the data from Roeland Bom (Bom, 2009), who labeled a small set of the accelerometer data for his MSc thesis.

This paper will present the reader with the problems we had with the data, the representation of the behaviours of the oystercatcher, the features we subtracted from the data, an attempt to use PCA for reducing the dimensions and the results. Finally a conclusion can be drawn.

2 Accelerometer

The accelerometer is a solar-powered device which records the acceleration the forward, sideway and down direction *surge, sway, heave* respectively, as shown in figure 1. The device weights approximately 12-18g. The device measures the acceleration at 20Hz for a three second intervals, this is not constant due to the fact that the device is solar-powered and can only measure if it has enough power. This results in a large variation of the number of measurements conducted each day. Which ranges from as low as 50 during winter season to 10000 during the summer season.

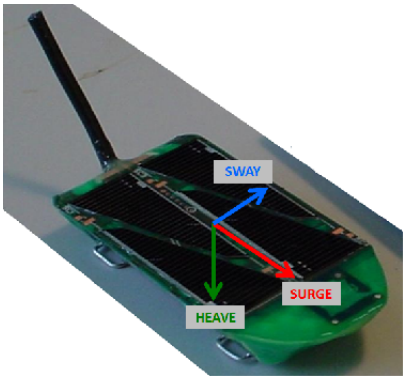


Figure 1: The Accelerometer

3 Data

3.1 Unlabeled Accelerometer data

The dataset to which we got access through W. Bouten is located on Sara. The dataset contains over 16 million accelerometer data from 57 birds. In the dataset acceleration is recorded from June 2009 till present. We could not use all the data and will elaborate on which data we used in the preprocessing section. We've created a MySQL database system so we could manipulate the data more freely. The data is saved in various tables in which the device id is linked to accelerometer data, time, speed and gps location data.

3.2 Labeled Accelerometer data

The dataset of the labeled data is also located on Sara. The data was for 3 birds which were manually labeled byBom (2009). The accelerometer data was measured from may till July 2009, which is also the breeding season. This data had static and dynamic accelerometer data instead of raw accelerometer data, which resulted in difficulties such as matching the data. Another difficulty was that Roeland's accelerometer data had a completely different scale than comparable data in the unsupervised dataset. A possible explanation for this is that he calibrated the acceleration data, something that was not possible with the unsupervised dataset because we did not have access to that information.

4 Representation of behaviours

In order to be able to understand the results of this research it is important to know how to interpret the acceleration data. Fortunately the three behaviour classes that we are trying to find all have distinct acceleration

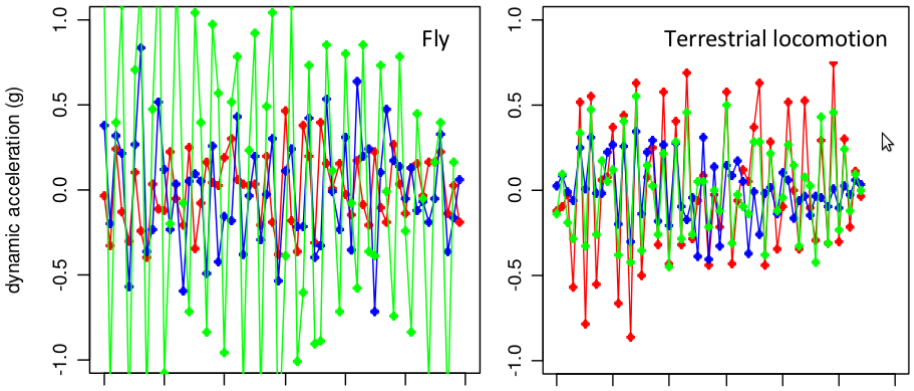


Figure 2: Standard picture of flying behaviour and terrestrial locomotion behaviour

graphs. The following sections describe the characteristics of the set of acceleration points measured over time.

4.1 Flying

Flying is the easiest behaviour to find, but is sadly rarely done. The Oysterchatcher only flies for about two percent of the time. Flying is so easy to separate because of its high variance in acceleration, see figure 2. This is caused by the flapping motion which causes the bird to go up and down and results in great acceleration differences.

4.2 No movement

No Movement is the opposite of flying and is characterized by very low variance and low acceleration in the x and y axis in general, see figure 3. The acceleration in the z axis can be higher if the bird is standing.

4.3 Terrestrial locomotion

Terrestrial locomotion behaviour is a little in between. There is quite some variance in the x and y axis and a bit less in the z axis.

4.4 Preprocessing

To make the data viable to use for classification, we had to remove many data from the data set. Data where the accelerometer was null, data where the longitude and latitude did not match with our knowledge of the Oystercatchers. The latter would be a problem because the speed of an oystercatcher

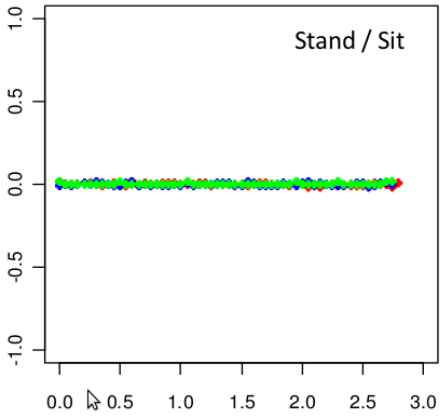


Figure 3: Standard picture of no-movement behaviour (sit and stand)

could well define its movement type, but if it went from Schiemonikoog to Hawaii in less than 4 hours is clearly a case of an outlier.

5 Features

When a measurement is started the device measures 20 times per second (20 Hz). These 20 points then are saved separately into the Flysafe database. One problem here is that one of these 20 points does not say anything about the behaviour during the second that the device was measuring. This means that we need to summerize the whole measurement into a number of features.

According to (Yang, Wang, & Chen, 2008) eight features, consisting of mean, correlation between axes, energy, interquartile range, mean absolute deviation, root mean square, standard deviation, and variance are usually extracted from triaxial acceleration data. We are going to use these eight features to summarize the whole measurement. Since the triaxial accelerometer collects signals from three axes, x-, y-, and z-axis, a total of 24 (3×8) features are calculated from a window of the acceleration data.

Most of the measurements from the Flysafe database are measurements with a duration of 3 seconds. Unfortunately not all the measurements have a duration of 3 seconds so we created new windows which consist of 20 measurements, for example if a measurement had a duration of 3 seconds (which means 60 measurement points) then we divided this measurement into three separate measurements of 20 measurement points. The reason for this is that in order to apply PCA to it later on we need every window to be of equal size. If you pick the window size too big then too many windows will be discarded because their duration is not long enough. By dividing every window into smaller windows of one second instead of three, you'll

make sure that in stead of discarding the entire window because its size is 2.9 seconds you only discard 0.9 second and keep two smaller windows of one second. You cannot make the window size too small, because then you won't have enough information to analyse. The features are then calculated for each window of one second.

6 Additional Features

6.1 Breading season

According to Bom (2009), Oystercatchers stay mostly close to their nests during the breading season, which is from May till July. This is either for feeding on the mudflats or for incubating and roosting on the saltmarsh. We thought that it would provide us some information to calculate whether the measurements of the accelerometer are measured in the breeding season of the Oystercatcher. Because we set up our own database we were able to query for the month in the date of the measurement and add to the data whether the month was in the breading season.

6.2 Territory

According to (Bom, 2009), the birds behaved differently inside their territory (less active) than outside their territory (more active). This means it would mean something to know whether the Oystercatcher is in his territory or not. So we added this also as an extra feature to the data.
It is said to be very easy to detect the territory of the bird because he spends a lot of time is his territory and when he is outside his territory it is far outside. So there should be a clear distinction between inside and outsite de territory. This enables us to simply determine the territory of the Oystercatcher.
Because the data provided us with the nest information of each Oystercatcher and the longitude and latitude at every measurement, we were able to calculate the distance to the birds nest using the haversin formula. After we calculated that automatically using MySQL we now knew how far the bird was from his nest.
To determine how big its territory is we made a plot which plots the frequency against the distance, shown in figure 4. We then determined manually what the bird his territory was. We came up with an average territory of 135 meters.

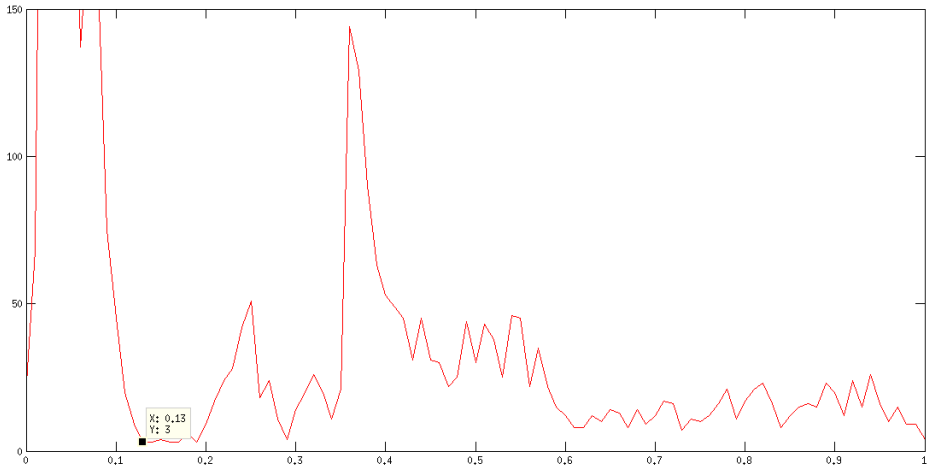


Figure 4: Frequency of distance to nest

7 Principle Component Analysis

7.1 Introduction

Principle Component Analysis is a quite useful data reduction and projection method. It is used in a lot of experiments to either reduce the amount of values per experiments so that comparison of the data for example in clustering can be done with fewer values or to center and rotate the axes of multidimensional measurements in the direction of the greatest variances with as few loss as possible. According to Alpaydin (2010) one trade-off however is a quite extensive calculation over the whole dataset, which can take exponentially longer with bigger datasets.

7.2 Application

We tried to use PCA as tool to reduce the amount of statistical features we calculated over the windows of measurements and hoped to get better results with more data, and at the same time requiring less time for clustering. To accomplish that we chose to reduce the 24 statistical features we extracted out of the windows to 8 PCA values which gave us a small enough relative error tolerance of 0.002
Also as recently found out PCA may be utilized to replace K-Means clustering (Ding & He, 2004) but we decided against it because we didn't want to rely only on data that is difficult to interpret (see 10.1.1) without even having the advantage of data reduction.

8 Unsupervised learning

The bird data that is acquired from the Flysafe database consist of accelerometer data. The problem with this data is that the data is not labeled with behaviours. However this is exactly what we are interested in, so we need to use a learning method which can find regularities in the input. An example of such a method is unsupervised learning.
A special case of unsupervised learning is called clustering. The aim of clustering is to find clusters or groupings of input (Alpaydin, 2010). We can use a clustering method to find clusters which can tell us something about the behaviour.

8.1 K-Means clustering

One clustering method is K-Means clustering. This method uses a distance measure to find clusters in the data. K-Means clustering is used because it is the most readily available technique and performs very well in most clustering tasks (Schreer, Hines, & Kovacs, 1998). This clustering method is an iterative procedure which is described by (Alpaydin, 2010) and works like this:

1. Start with some $\mathbf{m}_i$ initialized randomly
2. Use equation 1.1 to calculate b_i^t for all $\mathbf{x}^t$, which are the estimated labels
3. Set $\mathbf{m}_i$ to the mean of all the instances with equation 1.2
4. Repeat until $\mathbf{m}_i$ is stabilized

We use WEKA[1] which has an implementation of K-Means clustering available.

$$b_i^t = \begin{cases} 1 & \text{if } ||\mathbf{x}^t - \mathbf{m}_i|| = min_j ||\mathbf{x}^t - \mathbf{m}_j|| \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mathbf{m}_i = \frac{\Sigma_t b_i^t \mathbf{x}^t}{\Sigma_t b_i^t} \tag{2}$$

9 Clustering roadmap

The question here is how are we going to use clusters to know something about the behaviour of the oystercatcher? We can begin by saying that for every measurement point we now have 24 features. These features all say something about what the behaviour was in the *one-second measurement*.

[1] WEKA is available at `http://www.cs.waikato.ac.nz/ml/weka/`

We are going to use stepwise clustering to first separate flying behaviour from the rest, and then separate terrestrial locomotion behaviour from no-movement behaviour. This can be considered as a strategy to analyse the behaviour.

10 Results

10.1 Evaluation PCA

The results of clustering on the PCA values where not that good as with the statistical features itself, although we were able to find some characteristics that could prove useful when doing further research in clustering this data. We where able to extract fly almost as good as with the statistical features. Also we where able to see that behaviors with small movement attributes like sit stand and sleep can be clustered with moderate success in cluster2.(see table 1) Also it would be a supervised method because you actually would have to observe the bird first to analyze the data correctly in these situations, due to the disability to interpret PCA values (see section 10.1.1). Also for the tests we used re-sampling, which only is possible in supervised configurations.
We also compared the cluster results of one bird of the PCA values with a clustering of the same data with the statistical features and looked at the similarities inside the clusters. In other words we looked at how much of the data points in the clusters in PCA also could be found in the comparable cluster in the clusters of the statistical features. Because of the allocation of cluster1 and cluster3 we counted one time for aligned clusters on size and one time unaligned just by splitting alone. From the tables (see table 2) we see that the cluster we observed as fly can be extracted with almost the same accuracy from the PCA values as from the statistical features. The two biggest clusters seemed to be quite randomly allocated.

Table 1: PCA Clustering

0	1	2	3	4	5	assigned to cluster
60	44	0	43	42	30	Forage
34	41	0	48	50	42	Body care
10	38	0	27	24	126	Stand
54	54	0	11	61	29	Handle
15	20	0	56	40	68	sit
8	30	2	57	92	37	Aggression
19	52	0	39	57	44	Walk
0	23	0	43	5	125	Sleep

10.1.1 Problems

The main problem we encountered with the use of PCA was calculation time for the whole dataset of measurements. We tried different implementations of PCA lastly getting good results with a MATLAB package called qla (LLC, 2011). Qla is still under construction and is now released as a beta version. But the PCA part of the package worked very well. The self written PCA we tested was too slow to handle even medium amounts of data just as the PCA built-in from weka. They both used huge amounts of memory and computing power without even returning results after 2 days on 17000 measurements. The qla version was finally able to calculate the PCA values within seconds for the whole dataset of round 700.000 measurements.
Another problem with PCA is the interpretation of the data because one cannot interpret the values just as easily as looking at i.e. z-deviation.
Also because PCA is calculated over all measurements in the data if later on lots of outliers are detected, PCA has to be calculated again, because the outliers influence the results of the PCA calculation.

10.2 Stepwise clustering

10.2.1 Separate flying behaviour from the rest

The behaviour *fly* can easily be distinguished from the rest because this behaviour shows a high interquartile range in the z-axis compared to other types of behaviour. This is also shown in figure 2 where the z-axis is represented by the green line. So the first step we take is to cluster only on the interquartile range in the z-axis. We let WEKA create two clusters with K-Means clustering.
The result was very clear because K-Means clustering created two clusters, one with a mean interquartile range of 2160 and the other cluster with a mean interquartile range of 103.

10.2.2 Separate terrestrial locomotion from no-movement behaviour

The second step is to remove the cluster which we thought represented fly-behaviour and cluster again on the new data. Again we used K-Means

Table 2: PCA Comparison

	Statistical	PCA Unaligned	PCA Aligned	aligned	unaligned
cluster1	3427	2021	3063	1816	1384
cluster2	505	801	801	160	160
cluster3	1939	3063	2021	465	1071
cluster4	111	97	97	96	96
missed				3441	3267

clustering but this time let it create four clusters, this is done only on the interquartile range of x-,y- and z-axis. The method created four clusters, one of these clusters had values who were al below 50 and the other three were far above 50. We concluded that the cluster with the values below 50 is possibly a no-movement behaviour.

10.3 Evaluation of additional features

The adding of additonal features to cluster on did not show notable improvements in the clusters. Improvement in the sense that a cluster matches more to a specific behaviour model. It was actually often better for the cluster results to leave the additional features out. We had the intention to evaluate the additional features supervised instead by adding them to Roeland's data and see if the classification results would improve. Unfortunately Roeland only observed the Oystercatchers within the breeding season and within the territory.

10.4 Evaluation of the clustering roadmap

There are various ways to evaluate the clusters that were created by K-Means of which we chose three: Examining clustered points, looking at the resulting time budget and validating using a test set. In order to examine the clustered points we examined the clustered points by looking at the original window of acceleration data the point represents.

10.4.1 Examining Flying cluster

The original acceleration points that are plotted in figure 5 belong to a point in the *Fly* cluster. Looking back at the characteristics of flying we should see high amplitudes in the all signals, in particular in the z axis.

10.4.2 Examining No Movement cluster

The original acceleration points that are plotted in figure 6 belong to two points in the *No Movement* cluster. As is characteristic for behaviours grouped under the name of *No Movement* the variation is low, the acceleration graph typically approaches a straight line.

10.4.3 Examining Terrestrial locomotion cluster

The original acceleration points that are plotted in figure 7 belong to two points in the *Terrestrial locomotion* cluster. Characteristically *Terrestrial locomotion* has quite some variation in the x and y axes and a bit less variation in the z axis. It goes without saying that *Terrestrial locomotion* involves more movement than *No Movement* and less than *Fly*.

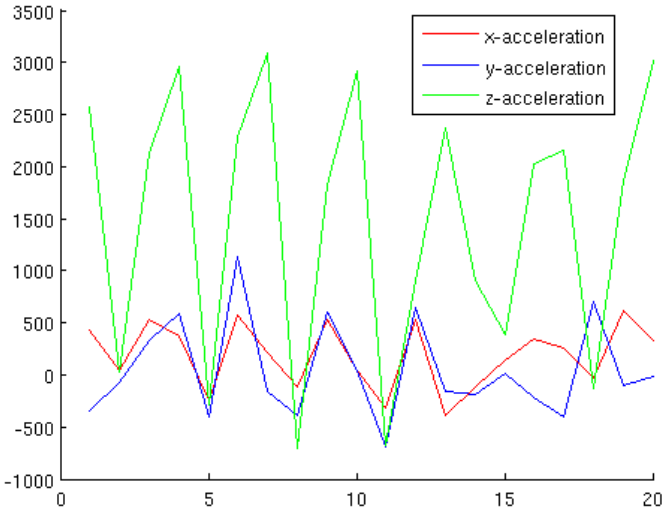


Figure 5: A window of acceleration points belonging to a point in the Fly cluster

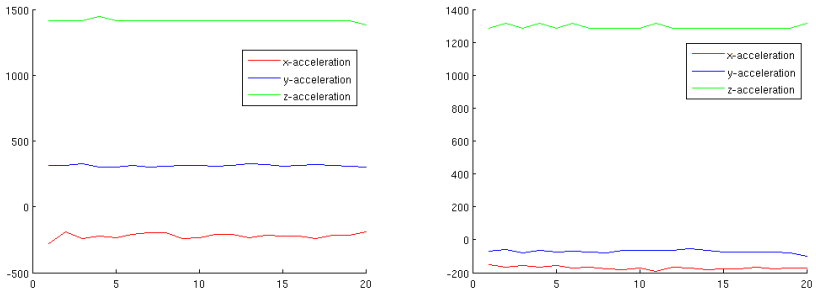


Figure 6: A window of acceleration points belonging to two points in the No Movement cluster

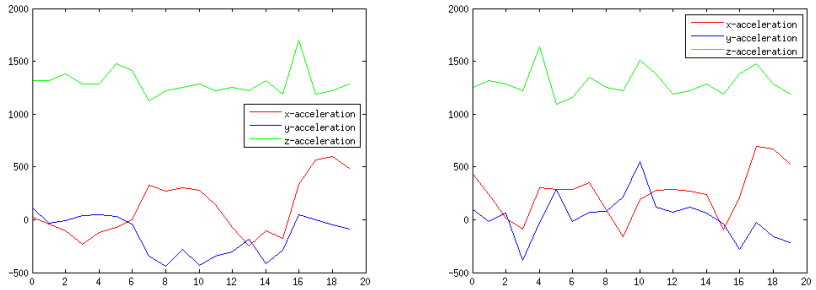


Figure 7: A window of acceleration points belonging to two points in the Terrestrial locomotion cluster

10.5 Examining Time Budget result

If we calculate the percentage of each cluster size we can analyse the amount of time that was spent on each type of behaviour. Actually wat is measured here is the percentage of datapoints that is part of a specific type of behaviour according to the clustering. The difference is that the time elapsed between each point is not regular. That means that behaviour that is more present in the breeding season will have a bigger influence than behaviour that is not.

Terrestrial locomotion	Flying:	No movement
39 %	2%	59 %

10.6 Evaluating clusters with a test-set

What we would like to do is now evaluate our own classified behaviours (fly, terrestrial locomotion and no-movement). To do this we took the data provided by Bom (2009) and calculated the 24 features (described earlier). Next we divided the set into a test-set, which was 18% of the whole set, and a training-set. We then removed the behaviour from the training-set and transformed the behaviours into the three behaviours (fly, terrestrial locomotion and no-movement) in the test-set.
We now applied the stepwise clustering method on the training-set to name the clusters as the behaviours; fly, terrestrial locomotion and no-movement. After that it was possible to test our manually classified behaviours with the test-set.

10.6.1 Classifying with a J48 tree

Now we have a training-set, with the manually named behaviours, and a test-set. Because the dataset of Bom (2009) was very little we decided to resample the data with the resample function in WEKA, we did this for the training-set and the test-set. The resample method in WEKA simply multiplies the existing data to create more data.
To classify the training-set we use the J48 tree learner which is implemented in WEKA. The result of this learning phase is a confusion matrix.

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
  39   0  16 |   a = fly
   0 137   5 |   b = no-movement
   0  33  50 |   c = terrestrial locomotion
```

As you can see 80 % of the data is correctly classified. But terrestrial locomotion is often classified as no-movement behaviour. Appendix A shows

you how the tree looks like.

11 Conclusion

We choose to summarize every measurement with eight features; mean, correlation between axes, energy, interquartile range, mean absolute deviation, root mean square, standard deviation, and variance. Some features were better in predicting the behaviour then others, especially the interquartile range was a good feature. PCA was used to reduce the 24 features into eight eigenvalues on which we used the unsupervised clustering method. The eigenvalues did not make an improvement in the clustering step, so we decided not to use PCA but to continue with another approach. The stepwise clustering approach pays off, because now we were able to separate first the flying behaviour from the rest and second, movement behaviour from no movement behaviour. The interquartile range in the z-direction appeared to be a very good predictor of the flying behaviour, if the interquartile range is bigger then a threshold of 1000 we could say that the behaviour is flying. To separate movement from no movement behaviour is well to be predicted with the interquartile range in the x-,y- and z-direction. Again a threshold can be considered but this time it has a value of 50. The results here are very promising, because now we can still know something about the behaviour of the Oystercatcher just by looking at the data. However it was very hard to distinguish in the groups we separated earlier. For example it was very hard to see if the bird is sitting or standing. However in the evaluation phase 80 % of the data was correctly classified. This means that the stepwise clustering method is a good method to start with naming behaviour in the data.

12 Further research

12.1 Future work

Future research for improving the results might be achieved by a collaboration between experts the biology field with experts on the machine learning field. Because a mutual understanding over birds should result in a better classification of the accelerometer data.

With experts in biology, who can classify animal behaviour different means of interpreting the accelerometer can be made. Such as how long a bird does a certain behaviour, with this information the terrestrial locomotion and no movement classification can be extended.

12.2 Improvements

Improvements for research in this area would be to acquire a better validation set, which can be achieved through collaboration with biology experts. Problems we had with the validation set were that they did not have the same format as the training set and that it was only in the breeding season and in the territory of the Oystercatcher. Having a validation set which covers this information would help adding more features which have a less statistical view on accelerometer data.

References

Alpaydin, E. (2010). *Introduction to machine learning* (Second ed.). London, England: The MIT Press.

Bom, R. A. (2009). *Can speed and tri-axial acceleration measured by biologgers be used to classify oystercatcher behaviour?* Unpublished master's thesis, University of Amsterdam.

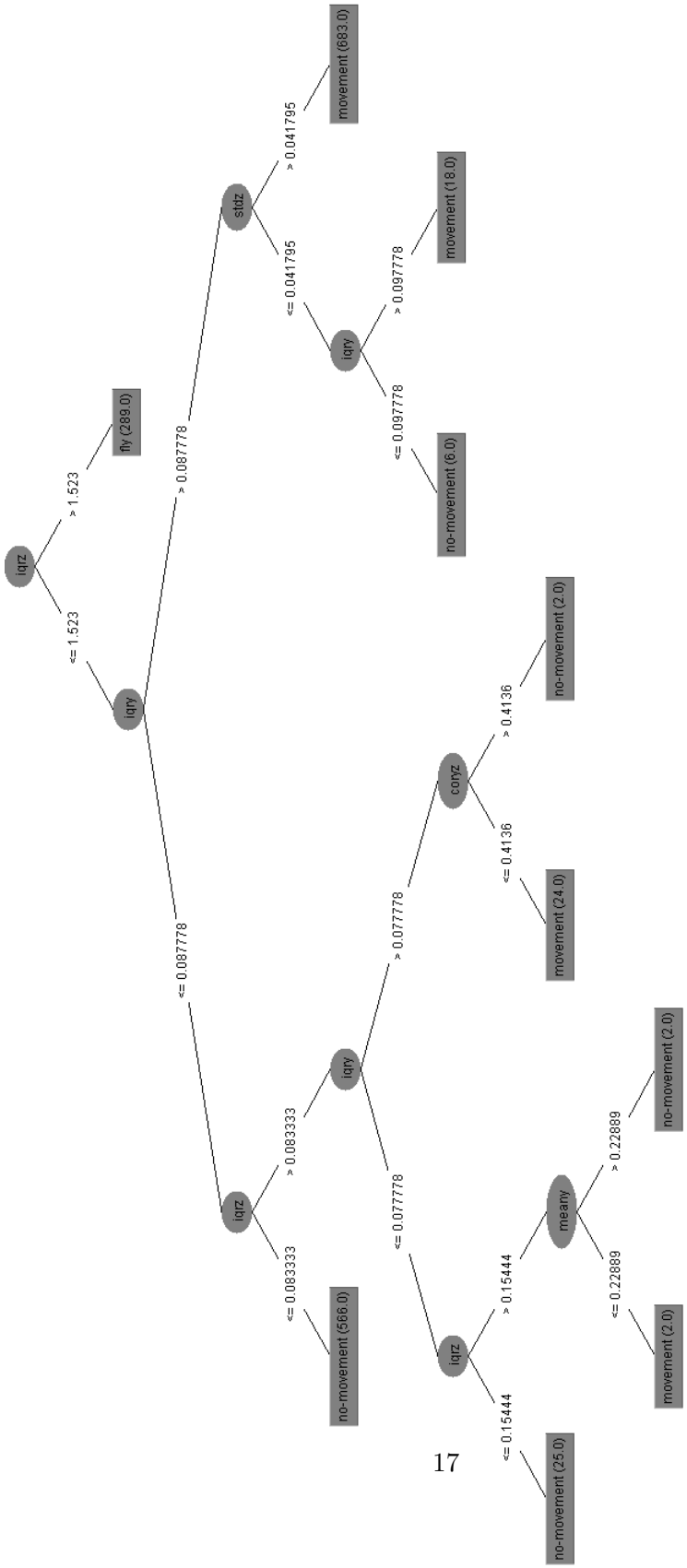
Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on machine learning* (p. 29).

LLC, M. A. (2011, jan). *The quick linear algebra library.* Available from `http://massiveanalytics.com/`

Schreer, J. F., Hines, R. J. O., & Kovacs, K. M. (1998). Classification of dive profiles: A comparison of statistical clustering techniques and unsupervised artifical neural networks. *Journal of Agricultural, Biological, and Environmental Statistics*, *3*(4), 383-404.

Yang, J.-Y., Wang, J.-S., & Chen, Y.-P. (2008). Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recognition Letters*, *29*(16), 2213 - 2220.

A Appendix A


iqrz
> 1.523
<= 1.523
fly (289.0)
iqry
> 0.087778
<= 0.087778
stdz
> 0.041795
<= 0.041795
movement (683.0)
iqry
> 0.097778
<= 0.097778
movement (18.0)
no-movement (6.0)
iqrz
> 0.083333
<= 0.083333
no-movement (566.0)
iqry
> 0.077778
<= 0.077778
conyz
> 0.4136
<= 0.4136
no-movement (2.0)
movement (24.0)
iqrz
> 0.15444
<= 0.15444
meany
> 0.22889
<= 0.22889
no-movement (2.0)
movement (2.0)
no-movement (25.0)