**CGI2015 manuscript No.**
(will be inserted by the editor)

# Supplementary Material - Evaluating TimeClassifier

**James S. Walker · Mark W. Jones · Robert S. Laramee · Owen R. Bidder · Hannah J. Williams · Rebecca Scott · Emily L. C. Shepard · Rory P. Wilson**

**Abstract** The following supplementary material provides additional evaluation and methods for the paper ”TimeClassifier - A Visual Analytic System for the Classification of Multi-Dimensional Time-Series Data”. Firstly we introduce the terms precision and recall which provide a quantitative measure for the success of classification algorithms in the data mining community. We then present the results of a user study with domain experts. Finally, we provide a comprehensive statistical comparison against traditional data mining approaches using these measures. These sections assist in the evaluation of our method to demonstrate its effectiveness for the analysis of animal behaviour.

**Keywords** Visual analytics · Time series analysis · Movement ecology

## 1 Precision and Recall

To gain a statistical measure of the effectiveness of our system, we utilize the information retrieval metrics, *time (T)*, *precision (P)* and *recall (R)* these are commonly used for evaluation in the data mining community. Precision ($P$) is the fraction of instances retrieved which are correct. While recall ($R$) is the fraction of correct instances successfully retrieved [1]. In terms of classification, a high recall means the algorithm classified most of the relevant results. A high precision corresponds to the algorithm correctly classifting a high proportion of the retrieved results. Precision and recall can be combined into one quantitative measure, called an F-Score ($F_1$). These are defined as:

$$P = \frac{|C|}{|B|} = \frac{|A| \cap |B|}{|B|} \quad R = \frac{|C|}{|A|} = \frac{|A \cap B|}{|A|} \quad F_1 = 2 * \frac{P * R}{P + R}$$

where given an input template, A is the set of correctly classified behaviour instances (taken from ground truth data), B is the set of retrieved entities (including those not correct), C is the set of correct classifications among the retrieved entities.

## 2 Domain Expert User Study

In addition to the formal timed user study with 30 participants on a subset of the data we also conducted a domain expert evaluation with 3 participants on the full data set. Whereas the formal user study used already segmented data and only needed a decision on whether each was to be labeled A or B, the domain expert study used the full unsegmented and unlabelled data set and thus shows a significant difference in the time required to manually label the data.

We obtained ground truth data from domain experts who analysed data from a deployment on a Penguin totalling 30 hours (864,319 data points recorded at 8 Hertz). Manually inspecting primarily the accelerometer attributes for patterns of Ascent, Descent, Burst Swim, and Swimming behaviours took approximately seven hours. Our software was augmented with logging capabilities to undertake an informal user study with three movement ecologists to identify these behaviours. Each participants precision and recall was logged over time (figure 1). On average, each participant obtained 97.3% precision and 88.9 % recall and took an average of 43 minutes to complete. It is probably that expert biologists in the field of marine wildlife users would be able to get closer to the 100% (as indicated in the field trials in the paper) due to their domain knowledge of the signals. Also the work with biologist experts indicated there is a learning effect using the program, such that it becomes faster with more familiarisation.

Rapid jumps occur in the precision and recall when our pattern searching was used. Slow increases and decreases indicate manual rejection and labeling. The time was notably split between locating templates and applying them to label the data. After applying our classification wizard, the users often went through the data rejecting misclassifications before manually adding missing results. Participant number one had prior knowledge of the data, and therefore produced the fastest and highest f-score. Participant two and three, had no prior knowledge of the data and therefore initially spent some time familiarizing themselves with the data. Interestingly, participant two experimented with different templates often using two or three templates for each behavior to increase the likely-hood of retrieving all the instances. Conversely, participant three opted to search for all the behaviors before rejecting / accepting results. They then applied additional templates if after visually inspecting the time-series the number of matches was not sufficiently high enough.

## 3 Evaluation Against Traditional Data Mining

We compare our visual analytic approach to several other traditional data mining approaches. These are, hierarchical clustering with DTW, 1NN with DTW, KNN, SVM, Random Forests and the Time Searcher 2 application. The Penguin data from the case study was further utilized to evaluate TimeClassifier against these methods. Table 1 displays results for precision and recall along with time for our approach without utilizing any interaction after the search. We applied a constant threshold throughout the pattern searching wizard with one template instance for each behavior. Near prefect scores are achieved for ascent and descent. Notable is the lower precision score for swimming and burst swimming, due to a low number of occurrences resulting in a low discriminating template.

| Behavior | (T) | (P) | (R) |
|---|---|---|---|
| Ascent | 3244 | 257 / 269 (96%) | 257 / 265 (97%) |
| Descent | 4741 | 259 / 277 (94%) | 259 / 265 (98%) |
| Swim | 5885 | 9 / 12 (75%) | 9 / 11 (82%) |
| Burst | 4975 | 21 / 25 (84%) | 21 / 24 (88%) |

Table 1: Table with time (in mili-seconds), precision, and recall results for data recorded from a Penguin

We gave each of the supervised learning algorithms 6 instances of each behavior, ranging from approximately 30 to 2000 data samples for each instance. The $x$, $y$, and $z$ axes of the accelerometer attribute form the feature vectors for each algorithm. The hierarchical clustering and 1NN algorithms required already segmented data, this is an unsolved problem in the time-series domain, however, we collected this from our ground truth data. Therefore those results are placed in italics and are only included for time and accuracy comparison and to show that even with ideal segmentation they would not perform as well as our approach. The results with respect to precision, recall and time are shown in table 2. The Matlab source code for these experiments is available in the supplementary material.

| Algorithm | (T) | (P) | (R) |
|---|---|---|---|
| *Hierarchical clustering with DTW* | 91 | * | * |
| *1NN with DTW* | 19 | 39% | 83% |
| KNN - X, Y, Z Feature Space | 7 | 22% | 56% |
| SVM - X, Y, Z Feature Space | 6871 | 21% | 71% |
| Random Forest - X, Y, Z Feature Space | 392 | 96% | 16% |
| Time Searcher 2 | 8 | 35% | 47% |
| Our System | 1.9 | 87% | 91% |

Table 2: Table comparing time (in seconds) along with average precision and recall scores for our TimeClassifier system results with state-of-the-art machine learning methods.

Results obtained show low precision and / or recall results for all existing classification techniques. High precision and recall scores are required to avoid a negative trade off. For example, random forests produces the highest precision of 96% but recall is low, only returning 16% of all behavior instances in the data. Conversely, 1NN returned 83% of all instances, however, only 19% were correct. Hierarchical clustering failed to return any meaningful results. We suspect the feature space results are low because they do not take into account the temporal ordering of variables, as such an overlap of values between templates increases misclassifications. Time searcher 2 is only capable of operating on a subset of 30,000 data items (3% of the original data size). Our solution produces robust precision and recall scores, higher than any of the existing methods utilized. Our results can be increased through interaction, and sufficient domain knowledge.

## References

1. Powers, D.M.W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Tech. Rep. SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)

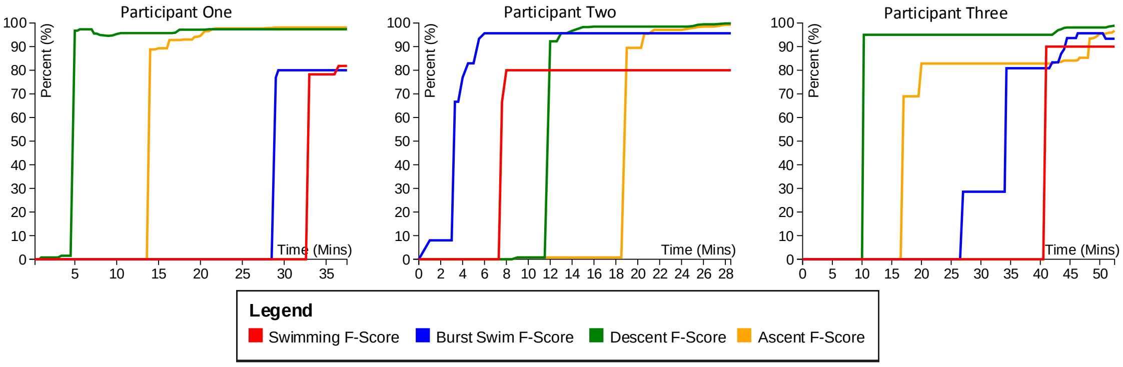

Fig. 1: This figure shows the three graphs for each of our participants. The $X$ axis encodes time, while the $Y$ axis encodes the F-Score percentage, a combination of precision and recall. Line color corresponds to the F-Score of the specified behavior detailed in the graph legend.