The IDA'01 Robot Data Challenge

Paul Cohen[1], Niall Adams[2], David J. Hand[2]

[1] Department of Computer Science. University of Massachusetts
cohen@cs.umass.edu
[2] Department of Mathematics. Imperial College, London
n.adams@ic.ac.uk, d.j.hand@ic.ac.uk

The IDA01 conference featured a *Data Analysis Challenge*, to which all conference participants could respond[3]. The challenge was organized around categorical time series data. A series of vectors of binary data was generated by the perceptual system of a mobile robot, and series of characters was taken from George Orwell's book *1984*. In both cases, the boundaries between meaningful units (activities in the robot data, words in the Orwell data) were absent, and part of the challenge involved inducing these boundaries.

More specifically, in each case we suspect a time series contains several patterns (where a pattern is a structure in the data that is observed, completely or partially, more than once) but we do not know the pattern boundaries, the number of patterns, or the structure of patterns. We suspect that at least some patterns are similar, but perhaps no two are identical. Finally, we suspect that patterns have a hierarchical structure in the sense that shorter patterns can be nested inside longer ones. The challenge is to find the patterns and elucidate their structure.

A *supervised* approach to the problem might involve learning to recognize patterns given known examples of patterns, however, this challenge encourages *unsupervised* solutions, those in which algorithms have no information specific to the data, other than the data itself. One reason for this stringent requirement is to see whether domain-general solutions will be developed. This is also the reason for providing two, quite different datasets: One hopes that an unsupervised pattern-finding algorithm that works on both data sets will provide some insights about general characteristics of patterns.

1 The Robot Dataset

The robot dataset is a time series of 22,535 binary vectors of length 9, generated by a mobile robot as it executed 48 replications of a simple approach-and-push plan. In each trial, the robot visually located an object, oriented to it, approached it rapidly for a while, slowed down to make contact, and attempted to push the object. In one block of trials, the robot was unable to push the object, so it stalled and backed up. In another block the robot pushed until the object bumped into the wall, at which point the robot stalled and backed up. In a third block of trials the robot pushed the object unimpeded for a while. Two trials in 48 were anomalous.

[3] Details of the challenge are available at http://genet.cs.umass.edu/dac/

Data from the robot's sensors were sampled at 10Hz and passed through a simple perceptual system that returned values for nine binary variables. These variables indicate the state of the robot and primitive perceptions of objects in its environment. They are: STOP, ROTATE-RIGHT, ROTATE-LEFT, MOVE-FORWARD, NEAR-OBJECT, PUSH, TOUCH, MOVE-BACKWARD, STALL. For example, the binary vector [0 1 0 1 1 0 1 0 0] describes a state in which the robot is rotating right while moving forward, near an object, touching it but not pushing it. Most of the $2^9 = 512$ possible states do not arise, in fact, only 35 unique states are observed. Fifteen of these states account for more than 97% of the time series. Said differently, more than half of the unique states occur very rarely, and five of them occur fewer than five times. Most of the 512 possible states are not semantically valid; for example, the robot cannot simultaneously be moving backward and moving forward. However, the robot's sensors are noisy and its perceptual system makes mistakes; for example, there are 55 instances of states in which the robot is simultaneously stalled and moving backward.

Because the robot collected ten data vectors every second, and its actions and environment did not change quickly, it is common to see long runs of identical states. The mean, median and standard deviation of run-length are 9.6, 4, and 15.58, respectively; while most runs are short, some are quite long.

Four forms of data are available:

1. 22,535 binary vectors of length 9.
2. 2345 binary vectors of length 9, produced by removing runs from dataset 1. The iterative rule for removing runs is: when two consecutive states are equal, keep the first and discard the second.
3. 22,535 numbers between 0 and 34. The original dataset contains only 35 unique vectors, so we can recode the vectors as numbers betwen 0 and 34, reducing the multivariate problem to a univariate one.
4. 2345 numbers between 0 and 34, obtained by recoding vectors as numbers and reducing runs.

The dataset was segmented into episodes by hand. Each of 48 episodes contained some or all of the following sub-episodes:

A: start a new episode with orientation and finding the target
B1: forward movement
B2: forward movement with turning or intruding periods of turning
C1: B1 + an object is detected by sonars
C2: B2 + an object is detected by sonars
D: robot is in contact with object (touching, pushing)
E: robot stalls, moves backwards or otherwise ends D

By hand, we associated one of these seven sub-episode type labels with each of the 22535 data items in the robot time series, producing an episode-labelled series of the same length.[4] The dataset contains 355 episodes.

A labelled subset of length 3558 of the original and univariate versions of the dataset is provided as part of the challenge, not to train supervised methods, but to test the performance of methods.

2 The Orwell Dataset

The task here is to take the first 5,000 words of George Orwell's *1984*, where spaces, capitalization and punctuation have been removed, and try to restore the word boundaries. We provide the dataset and also the locations of the word boundaries.

3 Results

At this writing, results are unavailable from IDA participants other than the authors of this report. Our qualitative conclusions are summarized here, and discussed in more detail in technical reports and papers in this volume.

What is a pattern? It is easy to write algorithms to look for structures in time series, but which structures should they look for? In a supervised approach, the answer is, "structures like those in the training data," but unsupervised algorithms must carry some bias to look for particular kinds of structures. Said differently, unsupervised pattern-finding algorithms define "pattern," more or less explicitly, as the sort of thing they find in data.

Most patterns are not meaningful. Patterns found by unsupervised pattern-finding algorithms are usually not meaningful in the domain to which they are applied. Said differently, for most conceptions of "pattern," there are many more patterns than meaningful patterns in a domain. To illustrate the point, consider the notion that patterns are the most frequent subsequences in a series. Listed from most to least frequent, here are the top 100 patterns in Orwell's text:

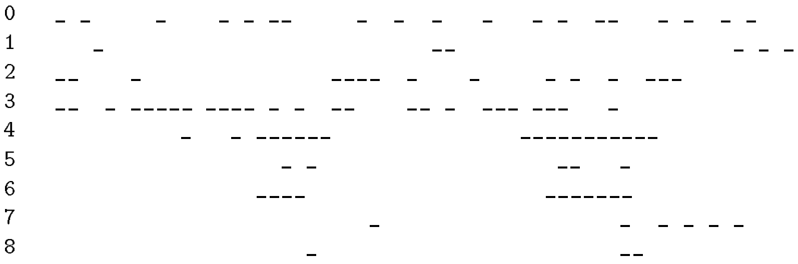
> th in the re an en as ed to ou it er of at ing was or st on ar and es ic el al om ad ac is wh le ow ld ly ere he wi ab im ver be for had ent itwas with ir win gh po se id ch ot ton ap str his ro li all et fr andthe ould min il ay un ut ur ve whic dow which si pl am ul res that were ethe wins not winston sh oo up ack ter ough from ce ag pos bl by tel ain

One sees immediately that most patterns (according to the frequency notion of pattern) are not morphemes in English; most are short, and the longer ones cross word boundaries (e.g., itwas, ethe, andthe). Clearly, if the patterns one seeks are English words or morphemes (or, as it happens, robot episodes) the notion that patterns are high-frequency subsequences is not sufficient. We can

[4] This cannot be done algorithmically, as some contextual interpretation of subsequences of the series is required. For example, if the sonars temporarily lose touch with an object, only to reaquire it a few seconds later, we label the intervening data C1 or C2, not B1 or B2, even though the data satisfy the criteria for B1 or B2.

load up our algorithms with bias to find domain-specific patterns, or try to develop a domain-general notion of pattern that has a better success rate than the frequency notion. The article by Cohen and Adams in this volume discusses the latter approach applied to the Challenge datasets.

Induction is necessary. Patterns often have variants and some kind of induction is required to generalize over them. The following image shows two episodes from the robot dataset (with runs removed). Each line represents an interval during which the corresponding value in the 9-vector was 1. The patterns are roughly similar in appearance, and, indeed, semantically similar; but they are not identical. They have different durations and somewhat different morphologies. A paper by Cohen in this volume describes a notion of pattern based on the temporal relationships between events that captures the essential structure of these data.



A session on the Challenge was held at the Intelligent Data Analysis symposium. Results from participants are summarized in technical reports and are available at the Challenge web site, `http://genet.cs.umass.edu/dac/`. There you will also find the Robot and Orwell datasets, as well as others. You are invited to try your methods and compare your results with other participants in the Challenge.

Acknowledgments

Prof. Cohen's work was supported by Visiting Fellowship Research Grant number GR/N24193 from the Engineering and Physical Sciences Research Council (UK). Additional support for Prof. Cohen and for Dr. Adams was provided by DARPA under contract(s) No.s DARPA/USASMDCDASG60-99-C-0074 and DARPA/AFRLF30602-00-1-0529.

This article was processed using the LaTeX macro package with LLNCS style