



Using sequenceMiner to Discover Anomalous Flights

SequenceMiner is an approach to model the behavior of discrete sensors in an aircraft during flights in order to discover atypical behavior of possible operational significance.

Background

We developed sequenceMiner to address the problem of detecting and describing anomalies in large sets of high-dimensional symbol sequence. The sequenceMiner can be used to discover atypical behavior that has possible operational significance for commercial or other flights. The sequenceMiner analyzes large repositories of discrete sequences and identifies operationally significant events. Our focus is on the primary sensors that record pilot actions. Each flight is analyzed as a sequence of events, taking into account both the frequency of occurrence of switches and the order in which switches change values. This method outperforms other leading technologies for sequence analysis in terms of speed, comprehensibility, and stability.

Research Overview

The approach taken performs unsupervised clustering (grouping) of sequences using the normalized longest common subsequence (LCS) as a similarity measure, followed by a detailed analysis of outliers to detect anomalies. Since LCS measure is expensive to compute, existing algorithms (methods) that have a high time-complexity, and often do not work well in practice. We present a new hybrid algorithm for computing the LCS that, in our tests, outperforms existing algorithms by a factor of five. We present new algorithms for outlier analysis that provide comprehensible indicators as to why a particular sequence was deemed to be an outlier. The algorithm provides a coherent description to an analyst of the anomalies in the sequence, compared to more 'normal' sequences. The algorithms we present are general and domain-independent, as there are also applications in related areas such as anomaly detection.

Figure 1: A Flight Path

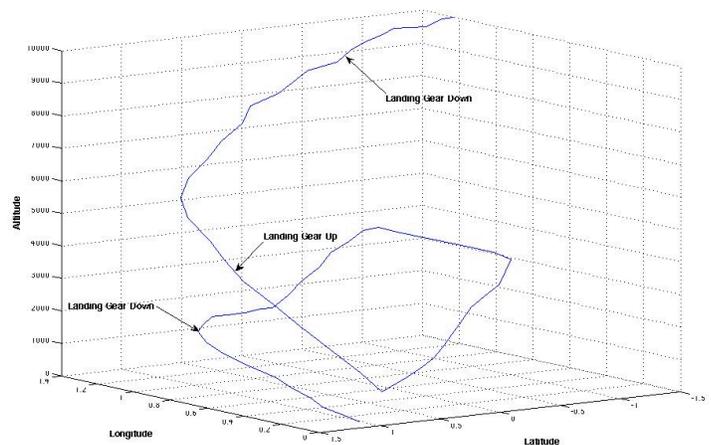
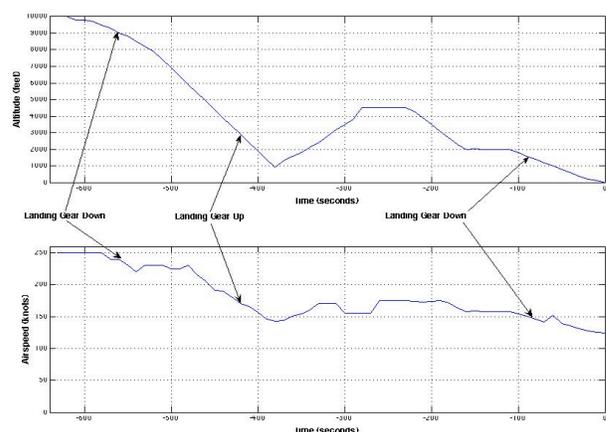


Figure 2: Flight altitude and airspeed graphs



How sequenceMiner works

Outline of Approach

Given a group of flights, the sequenceMiner should be able to find flights that are anomalous compared to the rest of the flights in the group. Given a flight known to be anomalous, sequenceMiner should be able to describe the anomalies in the flights, and the degree of anomalousness.

Approach to Task 1: Sequence Clustering

The sequenceMiner automatically find groups of flight sequences in data with high degree of similarity. We use the normalized longest common subsequence (nLCS) as the similarity measure for comparing flight sequences. The similarity measure takes into account the number of common switches, and also the order in which they are pressed. The clustering algorithm we use is CLARA (Clustering LARge Applications). This algorithm is a specially fast algorithm implemented to handle exceptionally large flights. CLARA automatically chooses the correct algorithm to use based on flight length. The algorithm can cluster 7,400 flights in 6 minutes.

Addressing Task 2: Identifying anomalous events in a flight

The sequenceMiner identifies sequences that have a significant number of switches that were pressed out of order, or were not pressed where expected, compared to the rest of the sequences. The two types of anomalous events are missing switches and excess switches. A missing switch occurs when a switch should have been pressed at a certain point in a flight, but was not. An excess switch occurs when a switch was pressed at a location where it was not expected.

Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment is a method used in bioinformatics to compare DNA sequences of organisms descended from a common ancestor is able to identify points of mutation inside a given sequence, by comparing it to other sequences. When this is taken the context of flights, these mutations are the points where a flight deviated from the norm. Each sequence is weighed differently. The weight is based on the distance of the sequence from the center of the cluster. Shorter sequences are given more weight, based on the assumption that shorter sequences will contain fewer non-essential switches. While constructing the alignment, an attempt is made to align more strongly with the more highly weighted sequences.

Incorporating Operational Information

Suppose the algorithm calculates that a switch should have been pressed at a certain time, but was not. It searches to see if the switch was pressed within one minute of the expected time. If the switch was pressed within a minute, it ignores the alarm. This step reduces alarms by around 30%.

Conclusions

We have described a system designed with the aim of detecting anomalies in discrete flight data. It does so by clustering flight data sequences using the normalized longest common subsequence (nLCS) as the similarity measure. As the nLCS is expensive to compute, previous methods typically have a higher runtime complexity, and do not work so well in practice. Our new hybrid algorithm is many times faster. This is an important contribution as the LCS is a commonly used algorithm in many areas and the running time of LCS algorithms is a frequent bottleneck. We presented algorithms based on a Bayesian model of a sequence clustering that detect anomalies inside sequences. In doing this, we move beyond what most current anomaly detection systems achieve, in not only predicting which sequences are anomalous, but by providing explanations as to why these particular sequences are anomalous. Our approach is general and not restricted in any way to a domain, and these algorithms can be of interest in other areas such as anomaly detection and event mining.

Points of Contact:

Ashok N. Srivastava Ph.D.
Principal Scientist and Group Leader, Intelligent Data Understanding Group.
Telephone: 650-604-2409
E-Mail: ashok@email.arc.nasa.gov
Web: <http://ti.arc.nasa.gov/people/ashok>

Group Web Page:

dataminng.arc.nasa.gov

