Aviation Data Mining

David Pagels

University of Minnesota, Morris

The Issue

January 31st, 2000 Puerto Vallarta, Mexico to Seattle, Washington

The Cause

"A loss of airplane pitch control resulting from the in-flight failure of the horizontal stabilizer trim system jackscrew assembly's acme nut threads. The thread failure was caused by excessive wear resulting from Alaska Airlines' insufficient lubrication of the jackscrew assembly"

Figure: The jackscrew with acme nut threads [5].

Figure: Alaska Airlines Flight 261 Memorial [3].

Outline

Legend

Multiple Kernel Learning	Hidden Markov Models & Hidden Semi-Markov Models	Text Classification
		

Background

- The Data
- Kernels
- Hidden Markov Models and Hidden Semi-Markov Models
- Natural Language Processing
- Types of Learning

Aviation Data

- Real Flight Recorder Data
- Synthetic Flight Recorder Data (generated by the flight simulator FlightGear)
- Aviation incident reports

Narrative: 1

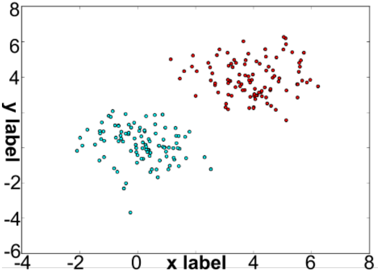
ON SHORT FINAL TWR TOLD ME TO GAR. I ACKNOWLEDGED AND PULLED UP GEAR IMMEDIATELY, TWR SAID 'DO A 360 DEG TURN TO THE R AND YOU'RE #1 TO LAND.' I THEN PUSHED GEAR CTL LEVER DOWN AND DID AS I WAS TOLD. R SEAT PAX SAID SHE HAD THE R WHEEL AND I VISUALLY CHKED L WHEEL, WHICH WAS DOWN, NO WARNING HORN. NEXT SOUND WAS THE SCRAPING OF THE BELLY ON THE RWY.

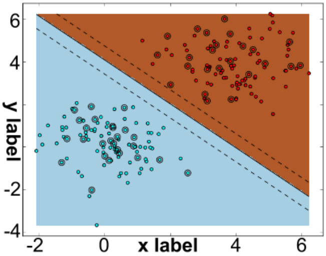
Kernels

Kernels

Similarity between vectors

Support Vector Machine

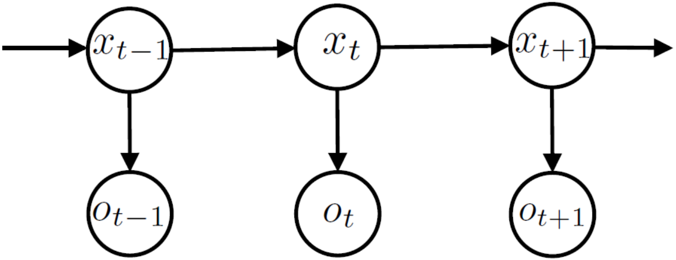




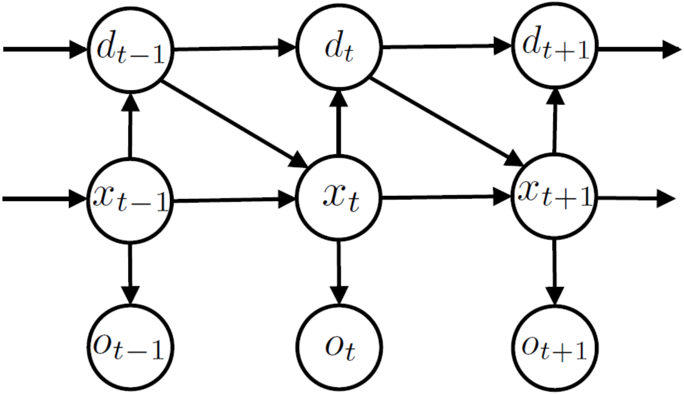
E. Kim. 2013

Hidden Markov Models and Hidden Semi-Markov Models

Hidden Markov Models



Hidden Semi-Markov Models



Natural Language Processing

Natural Language Processing

Extracting data from text generated by humans

Labels & text classification

Learning

- Supervised
- Semi-Supervised
- Unsupervised

Methods

The three methods

Multiple Kernel Learning

S. Das, B. L. Matthews, A. N. Srivastava, and N. C. Oza. 2010 [1]

The Problem

Heterogeneous Data: Discrete & Continuous

Compared to two baseline algorithms:

- Orca - Continuous
- SequenceMiner - Discrete

Longest Common Subsequence

Found using the Hunt-Szymanski Algorithm [2]

$$\vec{x}_i : ABB\,CBB\,AC$$

$$\vec{x}_j : AB\,A\,BA\,A\,C\,B$$

$$ABBAC$$

$$K_d(\vec{x}_i, \vec{x}_j) = \frac{5}{\sqrt{8 * 8}} = 0.625$$

Discrete Kernel

$$K_d(\overrightarrow{x}_i, \overrightarrow{x}_j) = \frac{5}{\sqrt{8 * 8}} = 0.625$$

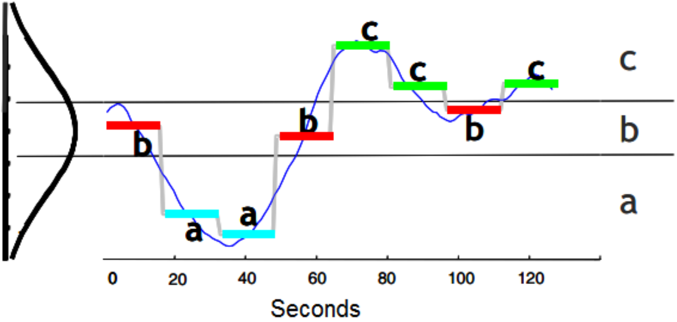
$$K_d(\overrightarrow{x}_i, \overrightarrow{x}_j) = \frac{|LCS(\overrightarrow{x}_i, \overrightarrow{x}_j)|}{\sqrt{l_{\overrightarrow{x}_i} l_{\overrightarrow{x}_j}}}$$

Continuous Kernel

Symbolic Aggregate approXimation (SAX) Representation

The same function as the discrete kernel.

SAX Representation



J. Lin, E. Keogh, L. Wei, and S. Lonardi. 2007

Combined Kernel

$$k(\overrightarrow{x}_i, \overrightarrow{x}_j) = nK_d(\overrightarrow{x}_i, \overrightarrow{x}_j) + (1-n)K_c(\overrightarrow{x}_i, \overrightarrow{x}_j)$$

Hidden Semi-Markov Models

Hidden Semi-Markov Model

I. Melnyk, P. Yadav, M. Steinbach, J. Srivastava, V. Kumar, and A. Banerjee. 2013 [4]

Normal Dataset

To find the probability of sequences, a set of 110 normal landings were generated using the flight simulator, FlightGear.

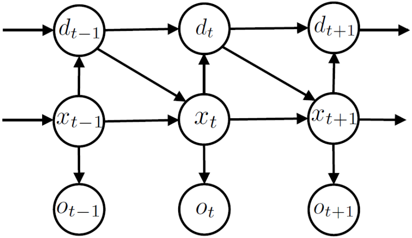
Anomalies

50 anomalies
10 of each:

1. Throttle is kept constant and flaps are not put down. The rest of the flight is the same as in normal case.
2. No initial throttle increase, the rest of the operation is normal.
3. The flight is similar to normal, except that the flaps are not put down.
4. At the end of the flight the brakes are not applied, the rest of the operation is normal.
5. Pilot overshoots the airport runway and lands somewhere behind it.

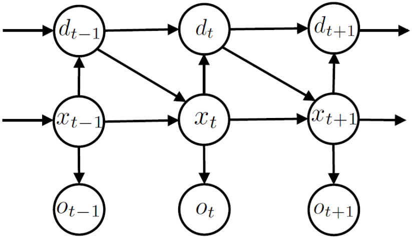
Sequence Probability

$$\frac{log\ \ p(o_1, o_2, \ldots, o_t)}{t}$$

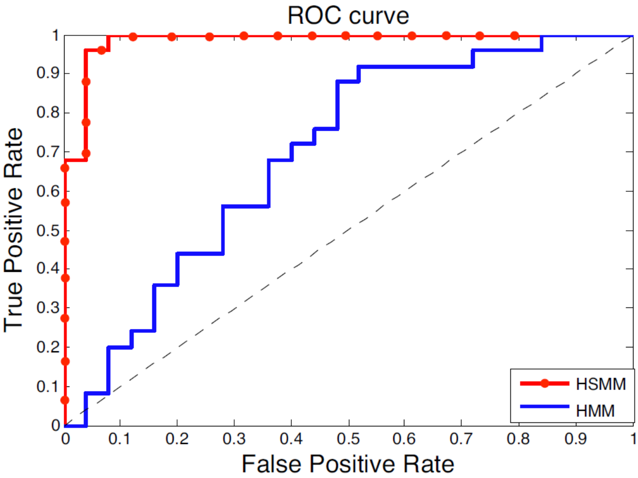


State Probability

$$p(o_t | o_1, o_2, \ldots, o_{t-1})$$



Receiving Operating Characteristic Curve



Text Classification

Classifying Aviation Incident Reports

I. Persing and V. Ng. 2009 [6]

Shapers and Expanders

Shapers are labels
Expanders indicate shapers
E.g. the expander 'snow' would indicate the 'Environment' shaper.

Shapers with Expanders

Shaping Factor	**Positive Expanders**	**Negative Expanders**
Physical Environment	cloud, snow, ice, wind	
Physical Factors	fatigue, tire, night, rest, hotel, awake, sleep, sick	declare, emergency, advisory, separation

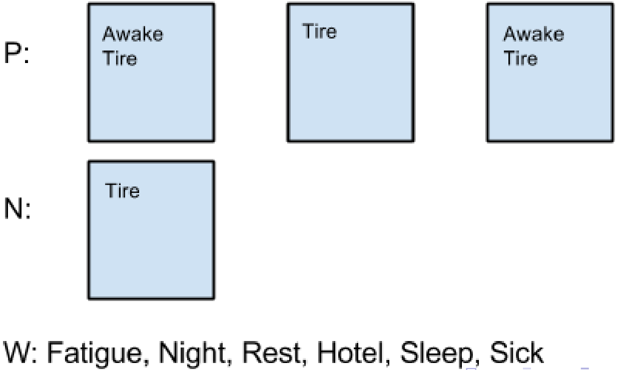
Bootstrapping Algorithm

- A set of positive examples of a shaper
- A set of negative examples of a shaper
- A set of unlabeled narratives
- Expand the largest set (positive or negative)
- Find 4 expanders

Finding the value for each word

Physical Factors shaper

$$t \leftarrow arg\ \ max_{t \notin W}(log\left(\frac{C(t,A)}{C(t,B)+1}\right))$$



Finding the maximum of those values

$$t \leftarrow arg\ max_{t \notin W}(log\left(\frac{C(t,A)}{C(t,B)+1}\right))$$

Tire: $log(\frac{3}{1+1}) = .176$

Awake: $log(\frac{2}{0+1}) = .301$

W: Fatigue, Night, Rest, Hotel, Sleep, Sick, Awake

Label Narratives

Assign shaper to narratives that contain ≥ 3 words in W

Results

Results of the three methods.

MKL Baseline Overlap

Algorithms	Overlap of anomalous flights (with MKAD)		
	Discrete	Continuous	Heterogeneous
O	21%	59%	34%
S	53%	0%	54%
O & S	58%	59%	67%
MKAD	**19**	**94**	**114**

Table: Overlap between MKAD approach and baselines. The baselines are represented by O for Orca and S for SequenceMiner. The values of O & S are the union of their anomalous sets [1].

HMM vs. HSMM

HSMM: Scenarios 1 and 2
Both: Scenarios 3, 4, and 5

1. Throttle is kept constant and flaps are not put down. The rest of the flight is the same as in normal case.
2. No initial throttle increase, the rest of the operation is normal.
3. The flight is similar to normal, except that the flaps are not put down.
4. At the end of the flight the brakes are not applied, the rest of the operation is normal.
5. Pilot overshoots the airport runway and lands somewhere behind it.

Text Classification Algorithm Comparison

Measured by a score composed of precision and recall.
Precision: Fraction of reports that were correctly labeled.
Recall: Fraction of reports that were correctly labeled out of the true number of reports that should have been labeled.
This score was 6.3% higher than the score from a purely supervised baseline [6]

Conclusion

Data mining techniques improving in aviation. We have discovered:

- How to detect heterogeneous anomalies more effectively
- HSMMs are better at detecting anomalies in aviation than HMMs
- A bootstrapping algorithm to find causes in aviation incident reports

Questions?

Resources I

S. Das, B. L. Matthews, A. N. Srivastava, and N. C. Oza.
Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study.
In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 47–56. ACM, 2010.

J. W. Hunt and T. G. Szymanski.
A fast algorithm for computing longest common subsequences.
In *Communications of the ACM: Volume 20-Number 5*, pages 350–353. ACM, 1997.

Resources II

D. Jenkins.
Sundial memorial to alaska airlines flight 261, port hueneme, california.
http://lost-at-sea-memorials.com/wp-content/uploads/2011/01/Mon1.jpg, 2011.

I. Melnyk, P. Yadav, M. Steinbach, J. Srivastava, V. Kumar, and A. Banerjee.
Detection of precursors to aviation safety incidents due to human factors.
In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 407–412. IEEE, 2013.

Resources III

NTSB.
Alaska airlines flight 261.
http://en.wikipedia.org/wiki/Alaska_Airlines_Flight_261#mediaviewer/File:Screwshavings2_sm.PNG, 2008.

I. Persing and V. Ng.
Semi-supervised cause identification from aviation safety reports.
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 843–851. Association for Computational Linguistics, 2009.