Interactive HMM construction based on interesting sequences

Szymon Jaroszewicz

National Institute of Telecommunications
Warsaw, Poland

LeGo 2008
Overview

- Building models interactively based on interesting patterns
- Hidden Markov Models
- Interesting patterns w.r.t. Hidden Markov Models
- Experimental evaluation: web server log
- Conclusions and Future research
Typical approach: Automatic model construction

Or:

Interactive HMM construction based on interesting sequences
Here: Interactive model construction

DATA → INTERESTING PATTERNS → USER

MODEL → MANUAL UPDATE

Szymon Jaroszewicz

Interactive HMM construction based on interesting sequences
Interactive model construction

- Understandable models
- Learn while building models
- Have to do ‘manual’ work :(
Previous related work

*Scalable pattern mining with Bayesian networks as background knowledge*
S. Jaroszewicz, T. Scheffer, D. Simovici
KDD’04, KDD’05, DMKD (to appear)

- Bayesian networks used as background model
- Exact and approximate algorithms given
- Models much closer to real relationships than automatically built models
Hidden Markov Models (HMMs)

Szymon Jaroszewicz

Interactive HMM construction based on interesting sequences
Hidden Markov Models (HMMs)

User gives the structure of the HMM:
- internal states
- which transitions are possible (not probabilities)
- which emission symbols are possible for each state (not probabilities)

Interactive HMM construction based on interesting sequences
Hidden Markov Models (HMMs)

User gives the structure of the HMM:

- internal states
- which transitions are possible (not probabilities)
- which emission symbols are possible for each state (not probabilities)

Diagram:

```
Sun ----> Rain
/     \
0.1 0.9
|     |
work↑<--play↓
/     \
0.8 0.2

Sun ----> Rain
```

Szymon Jaroszewicz

Interactive HMM construction based on interesting sequences
Interestingness of sequences w.r.t. an HMM

$$\text{Inter}(\text{seq}) = \left| \text{Prob}^{\text{HMM}}\{\text{seq}\} - \text{Prob}^{\text{Data}}\{\text{seq}\} \right|$$
Algorithm for finding all $\varepsilon$-interesting sequences

1. Train $HMM$ parameters based on $Data$ (Baum-Welch)
2. Find all $seq$ such that $\text{Prob}^{Data}(seq) > \varepsilon$
3. Find all $seq$ such that $\text{Prob}^{HMM}(seq) > \varepsilon$
4. Compute $\text{Prob}^{Data}$ for $seq$ frequent in $HMM$ but not in $Data$
5. Compute $\text{Prob}^{HMM}$ for $seq$ frequent in $Data$ but not in $HMM$
6. Compute $\text{Inter}(seq)$ for all sequences
7. Output $\varepsilon$-interesting sequences
Probability that sequence seq (starting at $t = 0$) is emitted and HMM ends in state $s_i$

$$\alpha(seq, s_i)$$

Efficient recursive updating:

$$\alpha(seq + o^{n+1}, s_i) = \sum_j \alpha(seq, s_j) P_{ji} E_{io^{n+1}}$$

$$\text{Prob}^{HMM}\{seq\} = \sum_i \alpha(seq, s_i)$$
Monotonicity property holds

\[ \text{Prob}^{HMM}\{seq + o\} \leq \text{Prob}^{HMM}\{seq\} \]

Standard depth-first frequent pattern mining works

*alpha* probabilities used instead of support counting

Very efficient: probability updating is fast
Web log format:


65.55.208.68 [01/Jan/2007:00:04:45] "GET /robots.txt" 200 51 "-" "msnbot/1.0"

Preprocessing:

- keep only top level directory
- sessionizing

Result: sessions:

journal/, journal/, __END__
robots.txt, index.html, journal/, ..., __END__
exchweb/, exchange/, exchange/, ..., __END__
...

Szymon Jaroszewicz
Interactive HMM construction based on interesting sequences
Initial HMM

all symbols \{\_\_END\_\_\}

1 \_all\_ 0.9182

0.0818 QUIT \

\_\_END\_\_
The Sophos antivirus

Top sequences:

- **sophos/,sophos/**
  - $\text{Prob}^{HMM} = 1.17\%$
  - $\text{Prob}^{Data} = 11.48\%$

- **sophos/,sophos/,sophos/,sophos/**
  - $\text{Prob}^{HMM} = 0.013\%$
  - $\text{Prob}^{Data} = 9.29\%$

- Update of the Sophos antivirus
- **Always** accessed 2, 4 or more times
The new model is:

- Each `soph` state only emits the `sophos/` symbol.
- `sophos/` symbol removed from `_all_` state.
Sequence: journals/, journals/, favicon.ico

\[ \text{Prob}^{HMM} \approx 0 \]
\[ \text{Prob}^{Data} \approx 2\% \]

- favicon.ico small icon next to web address

- Default location: main directory
  - At the Institute: img/ directory

- HTML header contains the other location; PDF can’t
- Browser tries the default location and fails
- Fixed: icon appears now
Added the following segment to the model:

The same PDF file often accessed twice; unable to explain:
- accelerators?
- browser errors?
- server errors?
Other patterns

- Exchange mail web reader
- robots: Google / MSN / Yahoo
- RSS readers
- ...

Szymon Jaroszewicz
Interactive HMM construction based on interesting sequences
Quickly built a model of high level user behavior

**Accuracy:** probability of all sequences modeled with error $< 0.01$

*Every* sequence is either:

- uninteresting (modeled well)
- infrequent

**Understandability:** the model is easily understandable

**Learnt** a lot about the data while modeling
Comparison with automatically learned models

- 20 hidden states + Baum Welch algorithm
- only transitions with prob. > 0.01
- all transitions with prob. > 0.001
Only transitions with prob. > 0.01
All transitions with prob. > 0.001

Szymon Jaroszewicz
Interactive HMM construction based on interesting sequences
Conclusions:

- Interactive model construction based on interesting patterns = **Understandability** + **Accuracy** + **Learning** about the data

Future work:

- Patterns starting at arbitrary time
- More general models: Dynamic Bayesian Networks, models of biological systems
- Automatic model updating (?)