Clustering Markov States into Equivalent Classes using SVD and Heuristic Search Algorithms

Xianping Ge, Sridevi Parise, Padhraic Smyth

Information and Computer Science
University of California, Irvine

`http://www.datalab.uci.edu/`

Abstract

Goal: To approximate a *large* Markov model (e.g., 10^5 states) with a *smaller* one (e.g., 50 states).

SVD-based Algorithm: Data matrix transformed to a (permuted) *block-constant* matrix. (Block $\equiv$ cluster)

Search-based Algorithms: Move a state from one cluster to another, so as to maximize data likelihood. (Search for "best" state $\rightarrow$ cluster assignment)

Too Many States?

- M-state Markov model has M^2 transition probabilities.
  - $M = 50000$ web pages on `www.ics.uci.edu`
  - $M \approx 400$ UNIX commands in Purdue UNIX user dataset
  - $M = 140072$ English words in *Wiretap/Classic* corpus
- Difficult to estimate the M^2 transition probabilities from limited data.
- Solution: cluster the states!

Problem Statement

Data: A set of sequences generated by a M-state Markov model.

- Or, equivalently, a $M \times M$ matrix of transition counts. ($n_{y,y'}$: number of times that y is followed by y'.)

Task: To cluster the M states into K clusters ("super-states"), where $K \ll M$.

- $\Rightarrow$ K-state Markov model

Goal: To maximize $P(\text{data}|K\text{-state Model})$.

SVD-Based Algorithm

- Permuted block-constant matrix: ($s(y)$: cluster of y)

$$H_{y,y'} \equiv P\Big(s(y')|s(y)\Big)/P\Big(s(y')\Big) \approx \frac{n_{y,y'}n}{n_y n_{y'}}.$$

$H_{y,y'}$ depends only on $s(y')$, $s(y)$ (i.e., not on y', y.)

s	1	1	2	2
1	a	a	b	b
1	a	a	b	b
2	c	c	d	d
2	c	c	d	d

s	1	2	2	1
1	a	b	b	a
2	c	d	d	c
2	c	d	d	c
1	a	b	b	a

- SVD of matrix with $K \times K$ constant blocks:
  - K nonzero singular values
  - Each singular vector is piecewise-constant w.r.t. the clusters.

SVD-Based Algorithm

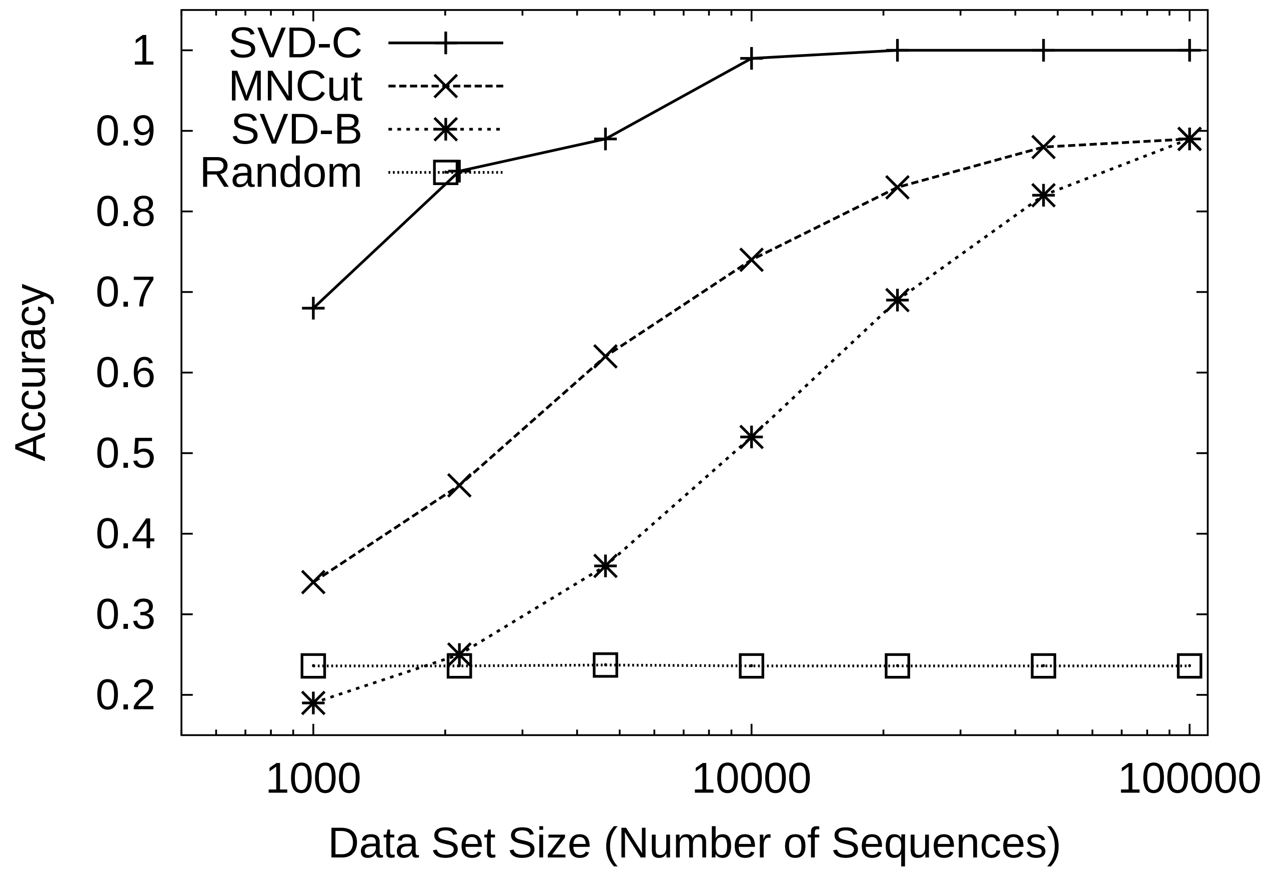
1. Run SVD on B where $B_{y,y'} = \frac{n_{y,y'} n}{n_y n_{y'}}$. Let the nonzero singular values be $\sigma_1, \ldots, \sigma_K$, corresponding to left and right singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K, \mathbf{v}_1, \ldots, \mathbf{v}_K$.

2. Run a clustering algorithm (e.g., K-means) on the rows of matrix

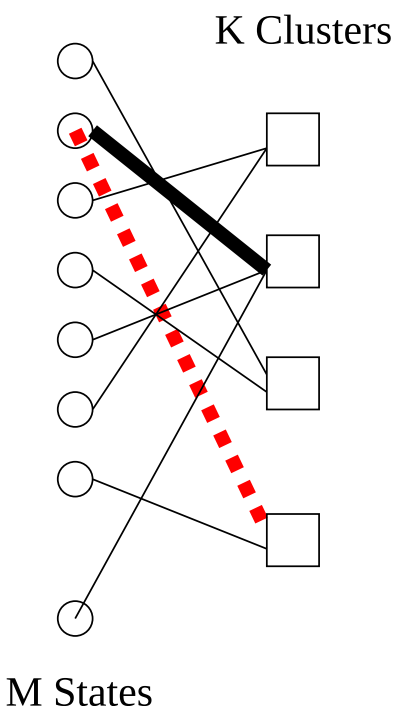
$$\left[\sigma_1 \mathbf{u}_1, \ldots, \sigma_K \mathbf{u}_K, \sigma_1 \mathbf{v}_1, \ldots, \sigma_K \mathbf{v}_K\right]$$

- SVD minimizes sum of squared errors (SSE) of matrix B which implies Gaussian noise: misfit with insufficient amount of data.
  - The above algorithm ("SVD-B") runs SVD on B where $B_{y,y'} = \frac{n_{y,y'} n}{n_y n_{y'}}$.
  - "SVD-C": run SVD on C where $C_{y,y'} = n_{y,y'}$.
- Related work: spectral clustering (e.g., in image segmentation)

Accuracy of the SVD methods



Clustering As a Search Problem



- Each clustering solution is an assignment of M states to K clusters.
- Search the solution space (of all possible assignments) for the best solution!
  - Start from a (random) initial assignment
  - Repeatedly try moving a state from one cluster to another, so as to increase data likelihood.

Search Algorithms

Score function : data likelihood

Heuristics :

GSAT: Of the $M \times (K-1)$ possible moves, choose the one that leads to the largest increase in the log-likelihood.

GSAT with 10% **sampling**

ICM: Go through the states sequentially, moving each state to the cluster that maximizes the log likelihood.

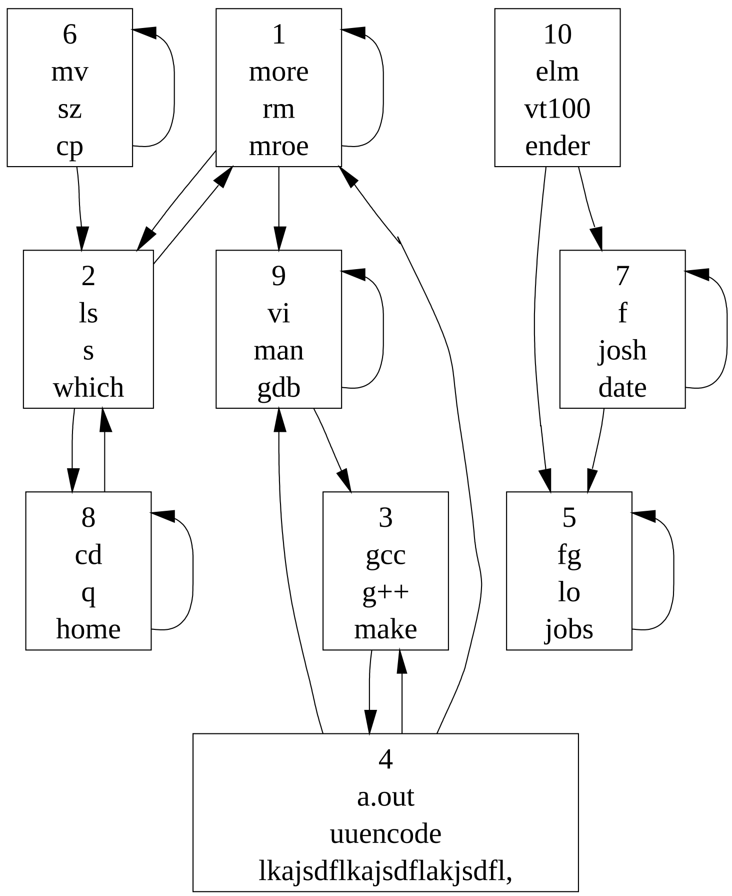
ICM with randomized order

Simulated Annealing

Accuracy & Computation Times on Simulated Data

Algorithm	Accuracy	Time/ICM
GSAT	0.889	16.8
GSAT with 10% sampling	0.881	3.3
ICM	0.885	1.0
ICM with randomized order	0.895	1.1
Simulated Annealing	0.959	316.3
SVD-C	0.681	18.1
Unconstrained HMM	0.43	1464.6
Random Assignment	0.234	-

Application to User Modeling: Purdue UNIX user data



- Data: Sequences of UNIX commands; each sequence is a session (from login to logout).
- $M \approx 400$ distinct UNIX commands.
- We ran ICM-style clustering algorithm on each user's sequences, with $K = 10$.
- Note the edit-compile-run cycle (states 9, 3, and 4).

Clustering of English Words

god	would	said	come
life	will	asked	go
night	could	cried	look
death	can	replied	love
course	did	says	use
water	should	answered	help
him	good	men	place
me	long	people	work
them	better	things	side
us	high	years	light
himself	true	words	part
home	dead	days	power

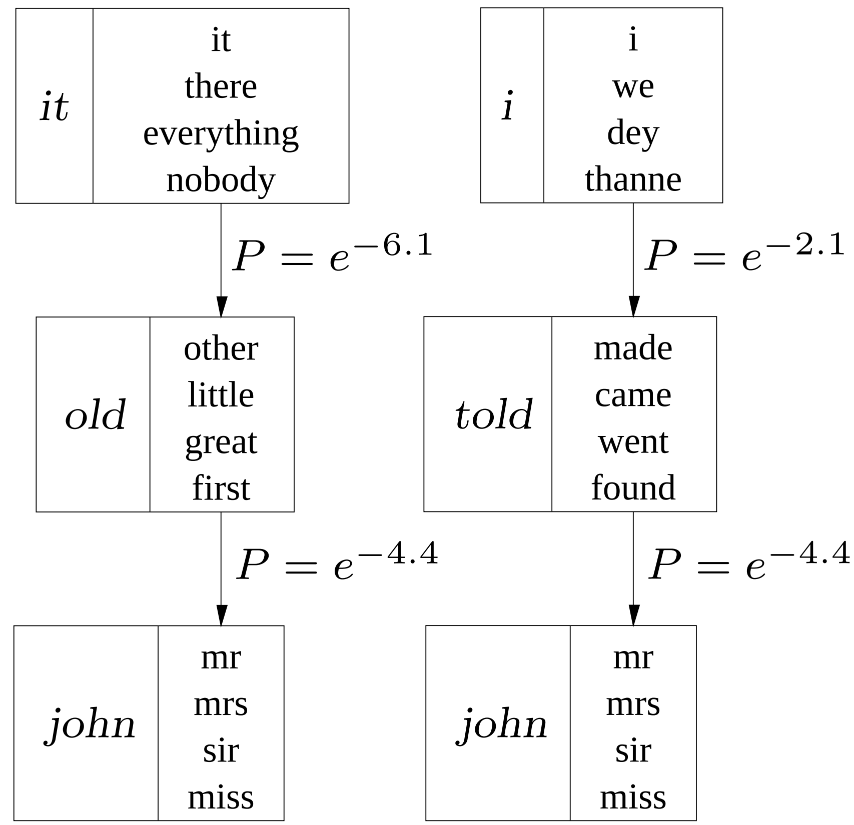
- Corpus: *Wiretap Online Library* / Classic. 92M bytes.
- Sequences: sentences.
- $M = 140072$ distinct words.
- $K = 48$ word clusters.
- See left for top words in 8 of the clusters.

Application to Word Segmentation

```
itoldjohnitwasyou
it old john it was you
i told john it was you

iamanamerican
i a man american
i am an american

whichendedinadeadheat
which ended in a dead he at
which ended in a dead heat
```



(a) (b)