Beating the Spread: Predicting Game Outcomes with a New Ranking Model.

Russell Albright
Anjela Govan
Amy Langville
Carl Meyer

data mining conference
M2008
Las Vegas

October 2008

Outline

1. Introduction
2. Offense-Defense Model
3. Other Ranking Methods
4. Data
5. Prediction Results

Basics of Ranking

- The *rank* of an object is its relative importance to the other objects in the finite set of size n. The ranks are 1,2,3, etc.
- Ranking models produce ratings.
- Ratings provide the degree of relative importance of each object.
- Applications of ranking include sports and search of web and literature.

ODM Development

A_{ij} = score team j generated against team i
$A_{ij} = 0$ otherwise

- *offensive rating* of team j

$$o_j = A_{1j}(1/d_1) + \ldots + A_{nj}(1/d_n)$$

- *defensive rating* of team i

$$d_i = A_{i1}(1/o_1) + \ldots + A_{in}(1/o_n)$$

$$\mathbf{o}^{(k)} = \mathbf{A}^T \frac{1}{\mathbf{d}^{(k-1)}}$$

$$\mathbf{d}^{(k)} = \mathbf{A} \frac{1}{\mathbf{o}^{(k)}}$$

Sinkhorn-Knopp Theorem (1967)

Definition

A square matrix $\mathbf{A} \geq 0$ is said to have total support if $\mathbf{A} \neq 0$ and if every positive element of $\mathbf{A}$ lies on a positive diagonal.

Theorem

For each $\mathbf{A} \geq 0$ with total support there exists a unique doubly stochastic matrix $\mathbf{S}$ of the form $\mathbf{RAC}$ where $\mathbf{R}$ and $\mathbf{C}$ are unique (up to a scalar multiplication) diagonal matrices with positive main diagonal.

A necessary and sufficient condition that the iterative process of alternatively normalizing the rows and columns of $\mathbf{A}$ will converge to a doubly stochastic limit is that $\mathbf{A}$ has support.

ODM convergence

- If $\mathbf{A}$ has total support $\rightarrow \{\mathbf{o}^{(k)}\}$, and $\{\mathbf{d}^{(k)}\}$ converge
- $\mathbf{A}$ may not have total support (but will have support)
- Can force total support

$$\mathbf{P} = \mathbf{A} + \epsilon \mathbf{e}\mathbf{e}^T$$

- As ϵ decreases number of iterations increases

ODM Algorithm

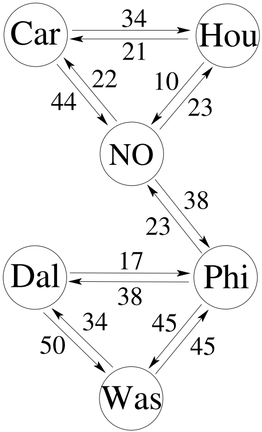
1. Represent the season using a weighted digraph with n nodes. On $i \rightarrow j$ the wight w_{ij} = amount of the statistic acquired by team j against team i.
2. Form adjacency matrix $\mathbf{A}$, $\mathbf{P} = \mathbf{A} + \epsilon \mathbf{e}\mathbf{e}^T$.
3. Team i has two rating scores, offensive o_i and defensive d_i

$$\mathbf{o}^{(k)} = \mathbf{P}^T \frac{1}{\mathbf{d}^{(k-1)}}$$

$$\mathbf{d}^{(k)} = \mathbf{P} \frac{1}{\mathbf{o}^{(k)}}$$

4. Overall rating score - rank aggregation (e.g. $r_i = o_i/d_i$).

2007 season NFL Example - ODM



Adjacency matrix $\mathbf{A}$:

	Car	Dal	Hou	NO	Phi	Was
Car	0	0	34	44	0	0
Dal	0	0	0	0	17	50
Hou	21	0	0	10	0	0
NO	22	0	23	0	38	0
Phi	0	38	0	0	0	45
Was	0	34	0	0	45	0

2007 season NFL Example (ODM)-result

- $\mathbf{A} + 0.001\mathbf{e}\mathbf{e}^T$, $tol = 0.01$

$$\mathbf{o} \approx \begin{pmatrix} 0.134 & 7.043 & 0.098 & 0.091 & 6.396 & 12.383 \end{pmatrix}^T$$

$$\mathbf{d} \approx \begin{pmatrix} 827.666 & 6.736 & 266.663 & 403.771 & 9.074 & 11.912 \end{pmatrix}^T$$

$$\mathbf{r} \approx \begin{pmatrix} 0.00016 & 1.0456 & 0.00037 & 0.00023 & 0.705 & 1.04 \end{pmatrix}^T$$

The list of ranked teams (from best to worst) is

Dal Was Phi Hou NO Car

Colley Method

1. Form Colley matrix $\mathbf{C}$

$$\mathbf{C}_{ij} = \begin{cases} -n_{ij} & \text{if} \quad i \neq j, \\ 2 + n_i & \text{if} \quad i = j, \end{cases}$$

   where n_i = total number of games played by team T_i and n_{ij} = number of times T_i played T_j.

2. Form vector $\mathbf{b}$

$$b_i = 1 + (w_i - l_i)/2,$$

   where w_i = number of T_i wins and l_i = number of T_i loses.

3. Solve

$$\mathbf{Cr} = \mathbf{b},$$

   the vector $\mathbf{r}$ contains rating scores of each team.

2007 season NFL Example - Colley Method

Car	16	NO	13
Dal	38	Phi	17
Dal	28	Was	23
Hou	34	Car	21
Hou	23	NO	10
NO	31	Car	6
Phi	33	Was	25
Phi	38	NO	23
Was	27	Dal	6
Was	20	Phi	12

Colley matrix $\mathbf{C}$:

$$
\begin{array}{c} \\ \text{Car} \\ \text{Dal} \\ \text{Hou} \\ \text{NO} \\ \text{Phi} \\ \text{Was} \end{array}
\begin{array}{c} \begin{array}{cccccc} \text{Car} & \text{Dal} & \text{Hou} & \text{NO} & \text{Phi} & \text{Was} \end{array} \\
\begin{pmatrix} 5 & 0 & -1 & -2 & 0 & 0 \\ 0 & 5 & 0 & 0 & -1 & -2 \\ -1 & 0 & 4 & -1 & 0 & 0 \\ -2 & 0 & -1 & 6 & -1 & 0 \\ 0 & -1 & 0 & -1 & 6 & -2 \\ 0 & -2 & 0 & 0 & -2 & 6 \end{pmatrix} \end{array}
$$

2007 season NFL Example (Colley)-result

$$\mathbf{r} \approx \begin{pmatrix} 0.3597 & 0.616 & 0.6687 & 0.3149 & 0.5015 & 0.5392 \end{pmatrix}^T$$

The list of ranked teams (from best to worst) is

Hou Dal Was Phi Car NO

Keener Method

1. Form Keener nonnegative matrix $\mathbf{K}$

- $$\mathbf{K}(i,j) = \begin{cases} h\left(\dfrac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right) & \text{team } i \text{ played team } j \\ 0 & \text{otherwise} \end{cases},$$

  where S_{ij} is the amount of points scored by team T_i against team T_j and

  $$h(x) = \frac{1}{2} + \frac{1}{2}\text{sgn}(x - \frac{1}{2})\sqrt{|2x-1|}$$

2. Rank vector $\mathbf{r}$ is the Perron vector of $\mathbf{A}$.

2007 season NFL Example - Keener Method

Car	16	NO	13
Dal	38	Phi	17
Dal	28	Was	23
Hou	34	Car	21
Hou	23	NO	10
NO	31	Car	6
Phi	33	Was	25
Phi	38	NO	23
Was	27	Dal	6
Was	20	Phi	12

Keener matrix $\mathbf{K}$:

	Car	Dal	Hou	NO	Phi	Was
Car	0	0	0.26	0.22	0	0
Dal	0	0	0	0	0.80	0.28
Hou	0.74	0	0	0.80	0	0
NO	0.78	0	0.20	0	0.26	
Phi	0	0.20	0	0.74	0	0.5
Was	0	0.72	0	0	0.5	0

2007 season NFL Example (Keener)-result

$$\mathbf{r} \approx \begin{pmatrix} 0.0474 & 0.2385 & 0.1107 & 0.1079 & 0.2342 & 0.2614 \end{pmatrix}^T$$

The list of ranked teams (from best to worst) is

Was Dal Phi Hou NO Car

Generalized Markov Method (GeM)

1. A sport season is a weighted directed graph with n nodes. Each game is loser $T_i \rightarrow$ winner T_j with weight w_{ij}= the positive difference of the game scores.
2. Form matrix $\mathbf{H}$

$$\mathbf{H}_{ij} = \begin{cases} w_{ij}/\sum_{k=1}^{n} w_{ik} & \text{if } i \text{ played } j \\ 0 & \text{otherwise} \end{cases}$$

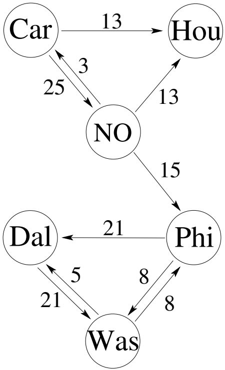
3. Form GeM matrix $\mathbf{G}$

$$\mathbf{G} = \alpha[\mathbf{H} + \mathbf{a}\mathbf{u}^T] + (1-\alpha)\mathbf{e}\mathbf{v}^T$$

   where $0 < \alpha < 1$, $\mathbf{v} > 0$ and $\mathbf{u}$ are probability distribution vectors and $a_i = 1$ if $\mathbf{H}_i^T = \mathbf{0}$ and 0 otherwise.
4. The vector containing the rating scores is $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{G}$$

2007 season NFL Example - GeM



$$\mathbf{H} + \mathbf{a}(1/6)\mathbf{e}^T =$$

	Car	Dal	Hou	NO	Phi	Was
Car	0	0	$\frac{13}{38}$	$\frac{25}{38}$	0	0
Dal	0	0	0	0	0	1
Hou	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
NO	$\frac{3}{31}$	0	$\frac{13}{31}$	0	$\frac{15}{31}$	0
Phi	0	$\frac{21}{29}$	0	0	0	$\frac{8}{29}$
Was	0	$\frac{5}{13}$	0	0	$\frac{8}{13}$	0

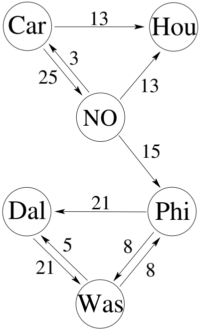
2007 season NFL Example (GeM)

$$\mathbf{G} = 0.85[\mathbf{H} + \mathbf{a}(1/6)\mathbf{e}^T] + 0.15(1/6)\mathbf{e}\mathbf{e}^T =$$

	Car	Dal	Hou	NO	Phi	Was
Car	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{6}{19}$	$\frac{111}{190}$	$\frac{1}{40}$	$\frac{1}{40}$
Dal	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{7}{8}$
Hou	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
NO	$\frac{133}{1240}$	$\frac{1}{40}$	$\frac{473}{1240}$	$\frac{1}{40}$	$\frac{541}{1240}$	$\frac{1}{40}$
Phi	$\frac{1}{40}$	$\frac{743}{1160}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{301}{1160}$
Was	$\frac{1}{40}$	$\frac{183}{520}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{57}{104}$	$\frac{1}{40}$

2007 season NFL Example (GeM)-result

$$\pi^T \approx \begin{pmatrix} 0.0389 & 0.2824 & 0.0656 & 0.056 & 0.2289 & 0.3281 \end{pmatrix}$$



The list of the teams in the order of rating scores (from best to worst) is

Was Dal Phi Hou NO Car

Point Spread

- Assume that
  point spread for game between T_i and T_j $=$
  $M|$rating $T_i -$ rating $T_j|$
- Use previous results to estimate M (Least Squares)

Data Gathering Challenges

- Reliable data sources
- Data format
- Amount of data
- Team names and league expansions

Data Gathering

- Sources - http://www.jt-sw.com/football/boxes/index.nsf (John M. Troan); http://scores.espn.go.com/ncf/scoreboard (ESPN);
- Data collection and parsing - automated with Perl scripts

NFL Game Prediction

- 2001-2007 with preseason padding
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM $\alpha = 0.6$

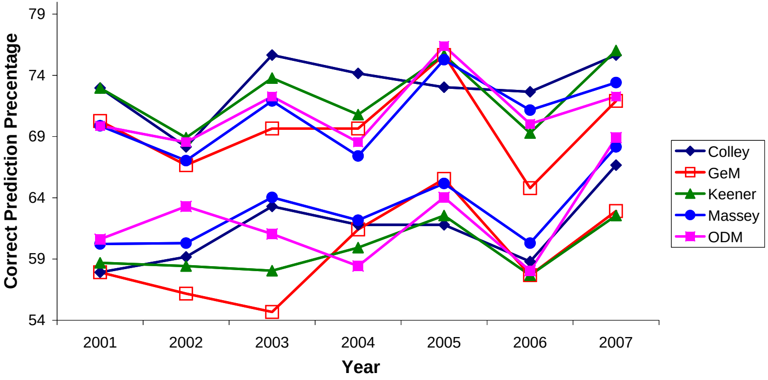
NFL Foresight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	57.92	57.92	58.69	60.23	60.62
2002	59.18	56.18	58.43	60.30	63.30
2003	63.30	54.68	58.05	64.04	61.05
2004	61.80	61.42	59.93	62.17	58.43
2005	61.80	65.54	62.55	65.17	64.04
2006	58.80	57.68	57.68	60.30	58.05
2007	66.67	62.92	62.55	68.16	68.91

NFL Hindsight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	72.97	70.27	72.97	69.88	69.88
2002	68.16	66.67	68.91	67.04	68.54
2003	75.66	69.66	73.78	71.91	72.28
2004	74.16	69.66	70.79	67.42	68.54
2005	73.03	75.66	75.66	75.28	76.40
2006	72.66	64.79	69.29	71.16	70.04
2007	75.66	71.91	76.03	73.41	72.28

NFL Foresight/Hindsight Prediction Results



NCAA Football Game Prediction

- Div I-A
- 2003-2007 starting week 5
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM $\alpha = 0.6$

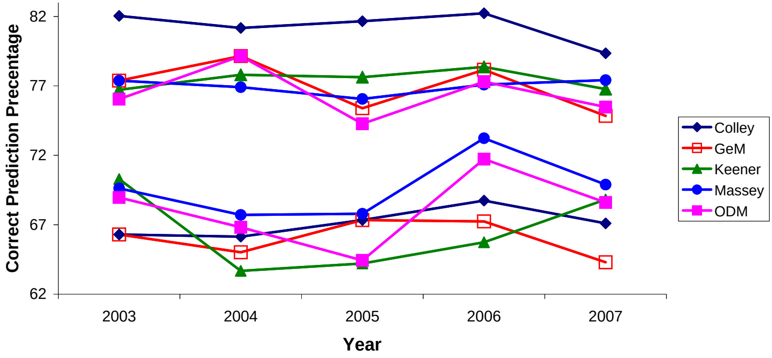
NCAA Football Foresight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2003	66.30	66.30	70.29	69.62	68.96
2004	66.14	65.02	63.68	67.71	66.82
2005	67.34	67.34	64.21	67.79	64.43
2006	68.74	67.24	65.74	73.23	71.73
2007	67.10	64.30	68.82	69.89	68.60

NCAA Football Hindsight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2003	82.04	77.38	76.72	77.38	76.05
2004	81.17	79.15	77.80	76.91	79.15
2005	81.66	75.39	77.63	76.06	74.27
2006	82.23	78.16	78.37	77.09	77.30
2007	79.35	74.84	76.77	77.42	75.48

NCAA Football Foresight/Hindsight Prediction Results



The End

Thank You! Questions?