The Child is Father of the Man: Foresee the Success at the Early Stage

Presenter: Liangyue Li (ASU)

Joint work by

Liangyue Li  
ASU

Hanghang Tong  
ASU
Important implications of high-impact scientific work:
- personal career development
- recruitment search
- jurisdiction of research resources

Question: how to forecast the long-term impact at the early stage?
Challenges

- C1: Scholarly feature design
- C2: Non-linearity
- C3: Domain heterogeneity
- C4: Dynamics
C1: Scholarly Feature Design

Obs: Adding content features brings little improvement
C2: Non-linearity

**Obs**: Non-linear methods outperform linear ones
**C3: Domain heterogeneity**

**Obs**: Impact of scientific work from different domains behaves differently.

- **pick up fast in early years**
- **Delayed pattern**
Question: How to quickly update the predictive model?
Roadmap

- Motivations
- **Proposed Solutions: iBall**
- Experimental Results
- Conclusions
iBall — Formulations

- Optimization Formulation

\[
\min_{w^{(i)}, i=1, \ldots, n_d} \sum_{i=1}^{n_d} \mathcal{L}[f(X^{(i)}, w^{(i)}), Y^{(i)}] + \lambda \sum_{i=1}^{n_d} \Omega(w^{(i)}) + \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} A_{ij} g(w^{(i)}, w^{(j)})
\]

- Remarks

- **Within-Domain Model**: regression/classification, linear/non-linear
- **Cross-Domain Consistency**: similar domains have similar models

**Question**: how to instantiate such consistency?
iBall — linear formulation

Details:

\[
\min_{\mathbf{w}^{(i)}, i=1,\ldots,n_d} \sum_{i=1}^{n_d} \left\| \mathbf{X}^{(i)} \mathbf{w}^{(i)} - \mathbf{Y}^{(i)} \right\|_2^2 + \lambda \sum_{i=1}^{n_d} \left\| \mathbf{w}^{(i)} \right\|_2^2 + \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \mathbf{A}_{ij} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|_2^2
\]

Intuitions: similar domain (large \( \mathbf{A}_{ij} \))

\( \rightarrow \) same feature has similar impact (small \( \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|_2 \))
iBall — non-linear formulation

Details:
\[
\min_{w^{(i)}, i=1,\ldots,n_d} \sum_{i=1}^{n_d} \| K^{(i)}w^{(i)} - Y^{(i)} \|_2^2 + \lambda \sum_{i=1}^{n_d} w^{(i)'}K^{(i)}w^{(i)} + \theta \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} A_{ij} \| K^{(i)}w^{(i)} - K^{(ij)}w^{(j)} \|_2^2
\]

Intuitions: similar domain (large $A_{ij}$)  
→ similar predicted outputs (small $\| K^{(i)}w^{(i)} - K^{(ij)}w^{(j)} \|_2^2$)
iBall — Closed-form Solutions

- Closed-form Solution

\[ w = S^{-1}Y \]

\[ \Rightarrow iBall — linear: \]

\[ w = [w^{(1)}; \ldots ; w^{(nd)}] \quad Y = [X^{(1)'}Y^{(1)}; \ldots ; X^{(nd)'}Y^{(nd)}] \]

\[ S = \begin{bmatrix} \vdots & \ldots & X^{(i)'}X^{(i)} + (\theta \sum_{j=1}^{nd} A_{ij} + \lambda)I & \ldots \\ \vdots & \ldots & -\theta A_{ij}I & \ldots \end{bmatrix} \]

Time Complexity: \( O((dn_d)^3) \)

\( d_n \): feature dim \( n_d \): # of domains

\( dn_d \) is in the order of 10 or 100
iBall — Closed-form Solutions

- Closed-form Solution

\[ w = S^{-1} Y \]

\[ w = [w^{(1)}; \ldots; w^{(n_d)}] \quad Y = [Y^{(1)}; \ldots; Y^{(n_d)}] \]

\[ S = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & (1 + \theta \sum_{j=1}^{n_d} A_{ij})K^{(i)} + \lambda I & -\theta A_{ij}K^{(ij)} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \]

Time Complexity: \( O(n^3) \)

\( n \): total # of training samples

\( n \) is in the order of millions
Key idea #1: Approx $S$ by low-rank approx

Details:

$$S_{t+1} \approx U_{t+1} \Lambda_{t+1} U'_{t+1}$$

(Overall: $O(n^2r)$)

$$w_{t+1} = S^{-1}_{t+1} Y_{t+1} = U_{t+1} \Lambda^{-1}_{t+1} U'_{t+1} Y_{t+1}$$

(Overall: $O(nr)$)

Complexity: $O(n^3) \rightarrow O(n^2r + nr)$

Benefit: avoid matrix inverse

Question: how to avoid re-computing low-rank approx at each time step?
**iBall — Scale-up with Dynamic Update**

- **Key idea #2:** Incrementally update the low rank structure of $S$

- **Details:**
  
  \[ S_{t+1} = \tilde{S}_t + \Delta S \]

  - **White:** zeros
  - **Blue:** old at $t$
  - **Pink:** new at $t+1$

- **Complexity:** \( O(n^2r) \rightarrow O((n + m)(r^2 + r'^2)), r \ll n \)

- **Benefit:** avoid re-computing low-rank approx

---

Roadmap

- Motivations
- Proposed Solutions: iBall
- Experimental Results
- Conclusions
Experiment Setup

- **Datasets**: AMiner\(^1\) (2,243,976 papers, 1,274,360 authors, 8,882 venues)
- **Evaluation Metric**: Root Mean Squared Error (RMSE)
- **Evaluation Objects**:
  - Effectiveness
  - Efficiency

\(^1\) [https://aminer.org/billboard/citation](https://aminer.org/billboard/citation)
Obs: iBall family joint models better than separate versions
Author Citation Prediction Performance

Obs: iBall family joint models better than separate versions
Venue Citation Prediction Performance

![Graph showing RMSE vs. Training Size for different models like iBall, iBall−fast, iBall−kernel, etc.]

Proposed Sol.

Obs: iBall family joint models better than separate versions
Error Analysis

Obs: bright region at x = y
Obs: iBall-fast outperforms other non-linear methods
Quality vs. Speed

Obs: iBall-fast: good trade-off between quality and speed
Roadmap

- Motivations
- Proposed Solutions: iBall
- Experimental Results
- Conclusions
Conclusions

- **Goals**: predict long-term impact of scholarly entities
- **Solutions**: joint predictive model (iBall)

<table>
<thead>
<tr>
<th>Challenges</th>
<th>C1 feature design</th>
<th>C2 non-linearity</th>
<th>C3 domain-heterogeneity</th>
<th>C4 dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tactics</strong></td>
<td>first 3 years’ citation</td>
<td>kernel trick</td>
<td>domain consistency</td>
<td>low-rank approximation</td>
</tr>
</tbody>
</table>

- **Results:**
  - iBall joint models better than separate versions
  - iBall-fast updates efficiently and accurately

- **More in paper:**
  - correctness and error bound analysis
  - significance and sensitivity tests