Feature Engineering and Classifier Ensemble for KDD Cup 2010

Chih-Jen Lin
Department of Computer Science
National Taiwan University

Joint work with HF Yu, HY Lo, HP Hsieh, JK Lou, T McKenzie, JW Chou, PH Chung, CH Ho, CF Chang, YH Wei, JY Weng, ES Yan, CW Chang, TT Kuo, YC Lo, PT Chang, C Po, CY Wang, YH Huang, CW Hung, YX Ruan, YS Lin, SD Lin and HT Lin
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
At National Taiwan University, we organized a course for KDD Cup 2010

Three instructors, two TAs, 19 students and one RA

19 students split to six sub-teams

Named by animals

Armyants, starfish, weka, trilobite, duck, sunfish

We will be happy to share experiences in running a course for competitions
Armyants

麥陶德 (Todd G. McKenzie), 羅經凱 (Jing-Kai Lou) and 解巽評 (Hsun-Ping Hsieh)
Starfish

Chia-Hua Ho (何家華), Po-Han Chung (鐘博翰), and Jung-Wei Chou (周融瑋)
Weka

Yin-Hsuan Wei (魏吟軒), En-Hsu Yen (嚴恩昱), Chun-Fu Chang (張淳富) and Jui-Yu Weng (翁睿妤)
Trilobite

Yi-Chen Lo (羅亦辰), Che-Wei Chang (張哲維) and Tsung-Ting Kuo (郭宗廷)
Duck

Chien-Yuan Wang (王建元), Chieh Po (柏傑), and Po-Tzu Chang (張博詞).
Sunfish

Yu-Xun Ruan (阮昱勳), Chen-Wei Hung (洪琛洧) and Yi-Hung Huang (黃曳弘)
Team Members

Tiger (RA)

Yu-Shi Lin (林育仕)
Snoopy (TAs)

Hsiang-Fu Yu (余相甫) and Hung-Yi Lo (駱宏毅)

Snoopy and Pikachu are IDs of our team in the final stage of the competition
Instructors

林智仁 (Chih-Jen Lin), 林軒田 (Hsuan-Tien Lin) and 林守德 (Shou-De Lin)
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Initial Thoughts and Our Approach

We suspected that this competition would be very different from past KDD Cups

- **Domain knowledge** seems to be extremely important for educational systems
- Temporal information may be crucial

At first, we explored a temporal approach

- We tried Bayesian networks
- But quickly found that using a **traditional** classification approach is easier
Initial Thoughts and Our Approach (Cont’d)

Traditional classification:

- Data points: independent Euclidean vectors
- Suitable features to reflect domain knowledge and temporal information

Domain knowledge, temporal information: important, but not as extremely important as we thought in the beginning
Our Framework

- Problem
  - Sparse Features
  - Condensed Features
  - Ensemble
Validation Sets

- Avoid overfitting the leader board
- Standard validation ⇒ ignore time series
- Our validation set: last problem of each unit in training set
- Simulate the procedure to construct testing sets

A unit of problems

\[
\begin{align*}
\text{problem 1} & \in V \\
\text{problem 2} & \in V \\
\vdots & \\
\text{last problem} & \in \tilde{V}
\end{align*}
\]

\(V\): internal training
\(\tilde{V}\): internal validation

In the early stage, we focused on validation sets
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
  - Condensed Features and Random Forest
  - Ensemble and Final Results
- Discussion and Conclusions
Sparse Features and Linear Classification

Problem

Sparse Features

Condensed Features

Ensemble
Basic Sparse Features

Categorical: expanded to binary features
- student, unit, section, problem, step, KC
Numerical: scaled by log(1 + x)
- opportunity value, problem view

A89: algebra_2008_2009
B89: bridge_to_algebra_2008_2009

<table>
<thead>
<tr>
<th>RMSE (leader board)</th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic sparse features</td>
<td>0.2895</td>
<td>0.2985</td>
</tr>
<tr>
<td>Best leader board</td>
<td>0.2759</td>
<td>0.2777</td>
</tr>
</tbody>
</table>

Five of six student sub-teams use variants of this approach

From this basic set, we add more features
Feature Combination and Temporal Information

- Feature combination: (problem, step) etc.
  $\Rightarrow$ Fetch hierarchical information
  Nonlinear mappings of data
- Temporal feature: add information in previous steps
  $\Rightarrow$ Fetch time series information
  e.g., add KC and step name in previous three steps as temporal features
Feature Combination and Temporal Information (Cont’d)

- **RMSE**
  - Basic
  - Basic + Temporal
  - Basic + Temporal + Combination
  - Basic + Temporal + More combination

- **Features**
  - A89
  - B89

- **Values**
  - 0.2985
  - 0.2883
  - 0.2875
  - 0.2836
  - 0.2816
  - 0.2815
Knowledge Component Feature

Originally using binary features to indicate if a KC appears. An alternative way:
Knowledge Component Feature

Originally using binary features to indicate if a KC appears. An alternative way:

Each token in KC as a feature

- “Write expression, positive one slope” similar to “Write expression, positive slope”
- Use “write,” “expression,” “positive” “slope,” and “one” as binary features
- Performs well on A89 only
Training via Linear Classification

- Large numbers of instances and features
- The largest number of features used is 30,971,151

<table>
<thead>
<tr>
<th></th>
<th>#instances</th>
<th>#features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A89</td>
<td>8,918,055</td>
<td>≥ 20M</td>
</tr>
<tr>
<td>B89</td>
<td>20,012,499</td>
<td>≥ 30M</td>
</tr>
</tbody>
</table>

- Impractical to use nonlinear classifiers
- Use LIBLINEAR developed at National Taiwan University (Fan et al., 2008)
- We consider logistic regression instead of SVM
- Training time: about 1 hour for 20M instances and 30M features (B89)
Leader board results:

<table>
<thead>
<tr>
<th></th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic sparse features</td>
<td>0.2895</td>
<td>0.2985</td>
</tr>
<tr>
<td>Best sparse features</td>
<td>0.2784</td>
<td>0.2830</td>
</tr>
<tr>
<td>Best leader board</td>
<td>0.2759</td>
<td>0.2777</td>
</tr>
</tbody>
</table>
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Problem

Sparse Features

Ensemble

Condensed Features
Condensed Features

Categorical feature ⇒ numerical feature

- Use correct first attempt rate (CFAR). Example: a student named sid:

\[
CFAR = \frac{\# \text{ steps with student} = \text{sid and CFA} = 1}{\# \text{ steps with student} = \text{sid}}
\]

- CFARs for student, step, KC, problem, (student, unit), (problem, step), (student, KC) and (student, problem)

Temporal features: the previous \( \leq 6 \) steps with the same student and KC

- An indicator for the existence of such steps
- Correct first attempt rate
- Average hint request rate
Temporal features:
- When was a step with the same student name and KC be seen?
- Binary features to model four levels:
  - Same day, 1-6 days, 7-30 days, > 30 days

Opportunity and problem view: scaled
Total 17 condensed features
Due to a small \# of features, we could try several classifiers via Weka (Hall et al., 2009).

Random Forest (Breiman, 2001) showed the best performance:

<table>
<thead>
<tr>
<th></th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic sparse features</td>
<td>0.2895</td>
<td>0.2985</td>
</tr>
<tr>
<td>Best sparse features</td>
<td>0.2784</td>
<td>0.2830</td>
</tr>
<tr>
<td>Best condensed features</td>
<td>0.2824</td>
<td>0.2847</td>
</tr>
<tr>
<td>Best leader board</td>
<td>0.2759</td>
<td>0.2777</td>
</tr>
</tbody>
</table>

This small feature set works well.
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Problem

Sparse Features

Condensed Features

Ensemble
Linear Regression for Ensemble

Linear regression to ensemble sub-team results

\[
\min_w \| y - Pw \|^2 + \frac{\lambda}{2} \| w \|^2
\]

- \( y \): labels of testing set: \( l \times 1 \); \( l \): \# testing data
- \( P \): \( l \times (\# \text{ results from students}) \)
- Truncated to \([0, 1]\): \( \min(1, \max(0, Pw)) \)
- Need some techniques as \( y \) unavailable

Decision of the regularization parameter \( \lambda \)
Ensemble Results

Ensemble significantly improves the results

<table>
<thead>
<tr>
<th></th>
<th>A89</th>
<th>B89</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic sparse features</td>
<td>0.2895</td>
<td>0.2985</td>
<td>0.2940</td>
</tr>
<tr>
<td>Best sparse features</td>
<td>0.2784</td>
<td>0.2830</td>
<td>0.2807</td>
</tr>
<tr>
<td>Best condensed features</td>
<td>0.2824</td>
<td>0.2847</td>
<td>0.2835</td>
</tr>
<tr>
<td><strong>Best ensemble</strong></td>
<td><strong>0.2756</strong></td>
<td><strong>0.2780</strong></td>
<td><strong>0.2768</strong></td>
</tr>
<tr>
<td>Best leader board</td>
<td>0.2759</td>
<td>0.2777</td>
<td>0.2768</td>
</tr>
</tbody>
</table>

- Our team ranked 2nd on the leader board
- Difference to the 1st is small; we hoped that our solution did not overfit leader board too much and might be better on the complete challenge set
## Final Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team name</th>
<th>Leader board</th>
<th>Cup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National Taiwan University</td>
<td>0.276803</td>
<td>0.272952</td>
</tr>
<tr>
<td>2</td>
<td>Zhang and Su</td>
<td>0.276790</td>
<td>0.273692</td>
</tr>
<tr>
<td>3</td>
<td>BigChaos @ KDD</td>
<td>0.279046</td>
<td>0.274556</td>
</tr>
<tr>
<td>4</td>
<td>Zach A. Pardos</td>
<td>0.279695</td>
<td>0.276590</td>
</tr>
<tr>
<td>5</td>
<td>Old Dogs With New Tricks</td>
<td>0.281163</td>
<td>0.277864</td>
</tr>
</tbody>
</table>

- Team names used during the competition:
  - Snoopy $\Rightarrow$ National Taiwan University
  - BbCc $\Rightarrow$ Zhang and Su
- Cup scores generally better than leader board
Outline

- Team Members
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Diversities in Learning

We believe that one key to our ensemble’s success is the diversity

- Feature diversity
- Classifier diversity

Different sub-teams try different ideas guided by their human intelligence
Diversities in Learning

We believe that one key to our ensemble’s success is the diversity

- Feature diversity
- Classifier diversity

Different sub-teams try different ideas guided by their human intelligence

Our student sub-teams even have biodiversity

- Mammals: snoopy, tiger
- Birds: weka, duck
- Insects: armyants, trilobite
- Marine animals: starfish, sunfish
Conclusions

- Feature engineering and classifier ensemble seem to be useful for educational data mining.
- All our team members worked very hard, but we are also a bit *lucky*.
- We thank the organizers for organizing this interesting and fruitful competition.
- We also thank National Taiwan University for providing a stimulating research environment.