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

- Introduction
- Course at NTU
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Outline

- Introduction
- Course at NTU
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
KDD Cup

- Annual data mining and knowledge discovery competition
- Organized by ACM special interest group on knowledge discovery and data mining
- 1997-present
- Now considered the most prestigious data mining competition
KDD Cup 2010

- Educational data mining competition
  https://pslcdata.datashop.web.cmu.edu/KDDCup/
- Predicting student algebraic problem performance given information regarding past performance
- Training data: summaries of the logs of student interaction with intelligent tutoring systems
- We refer to them as A89 and B89, respectively.
KDD Cup 2010 (Cont’d)

- Each data set: logs for a large number of interaction steps
- A89: 8,918,055 steps; B89: 20,012,499 steps
Log Fields

- student ID
- problem hierarchy including step name, problem name, unit name, section name
- knowledge components (KC) used in the problem
- number of times a problem has been viewed

Some log fields are only available in the training set:
- whether the student was correct on the first attempt for this step (CFA)
- number of hints requested (hint)
- step duration information.
Log Fields (Cont’d)

Hierarchy: step ⊂ problem ⊂ section ⊂ unit

Unit
CTA1_02 CTA1_01 ES_01 UNIT-CONVERSIONS-ONE-STEP

Section
CTA1_02-4 CTA1_01-4 ES_01-11 UNIT-CONVERSIONS-ONE-STEP-2

Problem
EG27 -5=-y PROP03 RATIO04-135 L2FB14B

Step
Series1AddPoint1 5=-y*(-1) ValidEquations R5C2
Log Fields (Cont’d)

KC examples:

KC subskills:
  Using simple numbers~~Find Y, any form~~Find Y
  Enter unit conversion
  Entering a given~~Enter given, reading words
  Entering a given~~Enter given, reading numerals

KC KTracedSkills:
  Identifying units-1
  Convert linear units-1~~Convert decimal units greater than one-1
  Select form of one with denominator of one-1
  Enter unit conversion-1
## Generation of Training/Testing Data

- **Testing data:** generated by randomly drawing a problem from a unit.
- **Problems before:** are used as training and after are discarded.

**A unit of problems**

<table>
<thead>
<tr>
<th>problem 1 $\in T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>problem 2 $\in T$</td>
</tr>
<tr>
<td>\vdots</td>
</tr>
<tr>
<td>problem $i$ $\in \tilde{T}$</td>
</tr>
<tr>
<td>problem $i+1$: not used</td>
</tr>
</tbody>
</table>

$T$: training \quad \tilde{T}$: testing
Competition Goal

- Predict CFA
- 0 (i.e., incorrect on the first attempt) or 1
- Training: CFA is available to participants
- A testing set of unknown CFA is left for evaluation
- Evaluation criterion: root mean squared error (RMSE)

\[ \sqrt{\frac{\|p - y\|^2}{l}} \]

- \( l \): # testing data, \( p \in [0, 1]^l \): predictions, \( y \in \{0, 1\}^l \): true answers
KDD Cup 2010 Schedule

- April 1: Registration opens at 2pm EDT, development data sets available
- April 19: Competition starts at 2pm EDT, challenge data sets available
- June 8: Competition ends at 11:59pm EDT
- June 14: Fact sheet and team composition info due by 11:59pm EDT
- June 21: Winners announced
- July 25: Workshop at ACM KDD 2010
Leaderboard

Based on results of a “unidentified” portion of testing data
Leaderboard (Cont’d)

Number of rows: 3925

<table>
<thead>
<tr>
<th>Overall Rank</th>
<th>Individual/Team Name</th>
<th>Algebra I 2008-2009</th>
<th>Bridge to Algebra 2008-2009</th>
<th>Total Score</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271157</td>
<td>0.272734</td>
<td>2010-06-08 23:46:36</td>
</tr>
<tr>
<td>2</td>
<td>NTU</td>
<td>0.274309</td>
<td>0.271162</td>
<td>0.272736</td>
<td>2010-06-08 13:28:24</td>
</tr>
<tr>
<td>3</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 22:30:56</td>
</tr>
<tr>
<td>4</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 22:32:22</td>
</tr>
<tr>
<td>5</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 13:09:07</td>
</tr>
<tr>
<td>6</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 22:25:29</td>
</tr>
<tr>
<td>7</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 13:43:21</td>
</tr>
<tr>
<td>8</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 22:18:28</td>
</tr>
<tr>
<td>9</td>
<td>NTU</td>
<td>0.274311</td>
<td>0.271163</td>
<td>0.272737</td>
<td>2010-06-08 22:22:56</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- **Course at NTU**
  - 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

Course page: http://www.csie.ntu.edu.tw/~cjlin/courses/dmcase2010/

Wiki: used to record progress
Team Members

- 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
- Every week each team reports progress
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 (嚴恩勗), 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)
Tiger (RA)

Yu-Shi Lin (林育仕)
Snoopy (TAs)

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

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

- Introduction
- Course at NTU
- **Initial Approaches and Some Settings**
  - Sparse Features and Linear Classification
  - Condensed Features and Random Forest
  - Ensemble and Final Results
- Discussion and Conclusions
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

![Diagram showing the relationship between Sparse Features, Condensed Features, and Ensemble in the framework.](image-url)
Validation Sets

- Avoid overfitting the leader board
- Standard validation \(\Rightarrow\) ignore time series
- Our validation set: last problem of each unit in training set
- Simulate the procedure to construct testing sets
- In the early stage, we focused on validation sets

\[
\begin{align*}
V & : \text{internal training} \\
\tilde{V} & : \text{internal validation}
\end{align*}
\]

A unit of problems

\[
\begin{align*}
\text{problem 1} & \in V \\
\text{problem 2} & \in V \\
\vdots \\
\text{last problem} & \in \tilde{V}
\end{align*}
\]
In the early stages, we focused on validation sets.

- Each sub-team submits to the leader board only once per week.

<table>
<thead>
<tr>
<th></th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal training</td>
<td>8,407,752</td>
<td>19,264,097</td>
</tr>
<tr>
<td>Internal validation</td>
<td>510,303</td>
<td>748,402</td>
</tr>
<tr>
<td>External training</td>
<td>8,918,055</td>
<td>20,012,499</td>
</tr>
<tr>
<td>External testing</td>
<td>508,913</td>
<td>756,387</td>
</tr>
</tbody>
</table>
This avoid overfitting the leaderboard

Of course in the end, many teams slightly violated the rule to submit more results in a week
Outline

- Introduction
- Course at NTU
- 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
- For example, 3,310 students in A89 ⇒ feature vector then contains 3,310 binary features to indicate the student who finished the step.

Numerical: scaled by $\log(1 + x)$
- opportunity value, problem view
- original range of opportunity in $[1, 1504]$, problem view in $[1, 18]$ for A89
- original range of opportunity in $[1, 2402]$, problem view in $[1, 29]$ for B89
- We have tried other scaling methods (e.g., linear scaling)
Basic Sparse Features (Cont’d)

A89: algebra_2008_2009
B89: bridge_to_algebra_2008_2009

<table>
<thead>
<tr>
<th>Data</th>
<th>stud.</th>
<th>unit</th>
<th>sec.</th>
<th>prob.</th>
<th>step</th>
<th>KC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A89</td>
<td>3,310</td>
<td>42</td>
<td>165</td>
<td>192,811×2</td>
<td>725,652</td>
<td>2,097×2</td>
</tr>
<tr>
<td>B89</td>
<td>6,043</td>
<td>50</td>
<td>186</td>
<td>53,375×2</td>
<td>129,349</td>
<td>1,699×2</td>
</tr>
</tbody>
</table>

- Number of features: 1M for A89, 200K for B89
- prob.: problem and problem view
- KC: KC and opportunity
Basic Sparse Features (Cont’d)

Results:

<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
Extensions from Basic Sparse Features

- Different scaling methods
- Slightly different ways to generate features
- Slightly different subsets of features
- Different regularization (L1 and L2) for classification

We will discuss some in detail
Due to large training size, nonlinear classifiers (e.g., kernel SVM) are not practical.

Linear classifier viable, but not exploiting possible feature dependence.

Following polynomial mapping in kernel methods or bigram/trigram in NLP, we use feature combinations to indicate relationships.

We manually identify some useful combinations for experiments.
Feature Combination (Cont’d)

- Example: hierarchical information
  (student name, unit name), (unit name, section name), (section name, problem name) and (problem name, step name)
- We have also explored combinations of higher-order features (i.e., more than two)
- We released two data sets using feature combinations at
  http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
- We thank Carnegie Learning and Datashop for allowing us to release them
Temporal Information

- Learning is a process of skill-improving over time.
- Temporal information should be taken into consideration.
- We considered a simple and common approach: For each step, step name and KC values from the previous few steps were added as features.
Feature Combination and Temporal Information

Leaderboard results

![Graph showing Leaderboard results with RMSE values and combinations](graph.png)

- Basic
- + Temporal
- + Combination
- + More combination

- A89
- B89

RMSE values:
- 0.2985
- 0.2883
- 0.2875
- 0.2836
- 0.2815
- 0.2816
Feature combinations very useful for B89
Temporal features more useful for A89
More features improve RMSE; but improvement less dramatic
Information already realized by earlier feature combinations
## Details of Features

<table>
<thead>
<tr>
<th>Combination</th>
<th>(student name, unit name), (unit name, section name), (section name, problem name), (problem name, step name), (student name, unit name, section name), (unit name, section name, problem name), (section name, problem name, step name), (student name, unit name, section name, problem name) and (unit name, section name, problem name, step name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>Given a student and a problem, add KCs and step name in each previous three steps as temporal features.</td>
</tr>
<tr>
<td>More combination</td>
<td>(student name, section name), (student name, problem name), (student name, step name), (student name, KC) and (student name, unit name, section name, problem name, step name)</td>
</tr>
</tbody>
</table>
## Number of Features

<table>
<thead>
<tr>
<th>Features</th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>1,118,985</td>
<td>245,776</td>
</tr>
<tr>
<td>+ Combination</td>
<td>6,569,589</td>
<td>4,083,376</td>
</tr>
<tr>
<td>+ Temporal</td>
<td>8,752,836</td>
<td>4,476,520</td>
</tr>
<tr>
<td>+ More combination</td>
<td>21,684,170</td>
<td>30,971,151</td>
</tr>
</tbody>
</table>
# Important Feature Combinations

<table>
<thead>
<tr>
<th>#features</th>
<th>A89</th>
<th>B89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>0.2895</td>
<td>0.2985</td>
</tr>
<tr>
<td>+ (problem name, step name)</td>
<td>0.2851</td>
<td>0.2941</td>
</tr>
<tr>
<td>+ (student name, unit name)</td>
<td>0.2881</td>
<td>0.2942</td>
</tr>
<tr>
<td>+ (problem name, step name) and (student name, unit name)</td>
<td>0.2842</td>
<td>0.2898</td>
</tr>
<tr>
<td>+ Combination</td>
<td>0.2843</td>
<td>0.2883</td>
</tr>
</tbody>
</table>

- (problem name, step name) and (student name, unit name) are very useful.
We tried many other ways
We will discuss some of them
They may be less effective than feature combinations mentioned earlier
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
Grouping Similar Names

- Two step names \(-18 + x = 15\) and \(5 + x = -39\) differ only in their numbers.
- For problem name and step name, we tried to group similar names together.
- By replacing numbers with a symbol, they become the same string and hence the same step name.
- Number of features reduced without deteriorating the performance.
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)
Logistic regression: CFA as label $y_i$

$$y_i = \begin{cases} 1 & \text{if CFA} = 1, \\ -1 & \text{if CFA} = 0, \end{cases}$$

Assume training set includes $(x_i, y_i), i = 1, \ldots, l$.

Logistic regression assumes the following probability model:

$$P(y \mid x) = \frac{1}{1 + \exp(-yw^T x)}.$$
Regularized logistic regression solves

\[
\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \log \left( 1 + e^{-y_i w^T x_i} \right) \quad (1)
\]

- \( w \): weight vector of the decision function,
- \( w^T w / 2 \): L2-regularization term, and
- \( C \): penalty parameter.

- \( C \): often decided by validation. We used \( C = 1 \) most of the time.
L2 regularization: a dense vector $\mathbf{w}$; we have also considered L1 regularization to obtain a sparse $\mathbf{w}$:

$$
\min_{\mathbf{w}} \|\mathbf{w}\|_1 + C \sum_{i=1}^{l} \log \left(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right) . \tag{2}
$$

Once $\mathbf{w}$ is obtained, we submitted either

$$
y = \begin{cases} 
1 & \text{if } \mathbf{w}^T \mathbf{x} \geq 0 \\
0 & \text{otherwise}
\end{cases}
$$

or probability values

$$
1/(1 + \exp(-\mathbf{w}^T \mathbf{x}))
$$
Using probability values gives a smaller RMSE than using 1/0

- Assume the true label is 0.
- Wrong prediction: errors using label/probability are 1 and $(1 - p_1)^2$
  
  $p_1 \geq 0.5$: predicted probability

- Correct prediction: errors are 0 and $p_2^2$, respectively.
  
  $p_2 \leq 0.5$: predicted probability
Quadratic function is increasing in $[0, 1]$,
Gain of reducing 1 to $(1 - p)^2$ is often larger than loss of increasing 0 to $p^2$.

Example:

<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th>error</th>
<th>label</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong</td>
<td>0.75</td>
<td>0.5625</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Correct</td>
<td>0.25</td>
<td>0.0625</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
We also checked linear support vector machine (SVM) solvers in LIBLINEAR. Result was slightly worse than logistic regression.
**Result Using Sparse Features**

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

- Introduction
- Course at NTU
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Condensed Features and Random Forest

Problem

Sparse Features

Condensed Features

Ensemble
Condensed Features

Categorical feature $\Rightarrow$ numerical feature

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

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

- CFARs for student, step, KC, problem, (student, unit), (problem, step), (student, KC) and (student, problem)
Condensed Features (Cont’d)

Temporal features: the previous ≤ 6 steps with the same student and KC
- An indicator for the existence of such steps
- Average of CFAs
- Average hints (up to six depending on the availability)

Other 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
Condensed Features (Cont’d)

Opportunity and problem view:
- First scaled by
  \[
  \frac{x}{x + 1}
  \]
- Then linearly scaled to \([0, 1]\)

Total 17 condensed features
- Eight CFARs
- Seven temporal features
- Two scaled numerical features for opportunity and problem view.
Training by Random Forest

- Due to a small number of features, we could try several classifiers via Weka (Hall et al., 2009).
- To save training time, we considered a subset of training data and split the classification task into several independent sets according to unit name.
- That is, for each unit name, we collected the last problem of each unit to form its training set.
- In testing, we checked the testing point’s unit name to know which model to use.
Random Forest (Breiman, 2001) showed the best performance:

10 decision trees with depth 7

<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><strong>Best condensed features</strong></td>
<td><strong>0.2824</strong></td>
<td><strong>0.2847</strong></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

Due to the small feature size, a Random Forest on the training subset of a unit takes a few minutes.
Outline

- Introduction
- Course at NTU
- Initial Approaches and Some Settings
- Sparse Features and Linear Classification
- Condensed Features and Random Forest
- Ensemble and Final Results
- Discussion and Conclusions
Ensemble and Final Results

Problem

Sparse Features

Condensed Features

Ensemble

Chih-Jen Lin (National Taiwan Univ.)
Past competitions (e.g., Netflix Prize) showed ensemble of results from different methods often boost the performance.

We find a weight vector to linearly combine predicted probabilities from student sub-teams.

We did not use a nonlinear way because a complex ensemble may cause overfitting.
Linear Regression for Ensemble (Cont’d)

- We checked linear models
  simple averaging, linear SVM, linear regression, logistic regression
- Linear regression gives best leaderboard result
  Probably because linear regression minimizes RMSE
  (the evaluation criterion)
Linear Regression for Ensemble (Cont’d)

- Given \( l \) testing steps and \( k \) prediction probabilities \( p_i \in [0, 1]^l, \quad i = 1, \ldots, k, \)

\[
\min_{\mathbf{w}} \quad \|\mathbf{y} - P\mathbf{w}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{3}
\]

\( \lambda \): regularization parameter, \( \mathbf{y} \): CFA vector, and \( P = [p_1, \ldots, p_k] \).

- If \( \lambda = 0 \), Eq. (3) just a standard least-square problem
In SVM or logistic regression, we may add a bias term $b$ so
\[ Pw \Rightarrow Pw + b1 \]
where $1 = [1, \ldots, 1]^T$.
We also replaced $\|w\|^2$ with $\|w\|^2 + b^2$.
The obtained weight $w$ is used to calculate $Pw$ for combining prediction results.
$Pw$ may be out of the interval $[0, 1]$. We employ a simple truncation:
\[ \min(1, \max(0, Pw)) \],
(4)

1: vector with all ones; 0: vector with all zeros.
We also explored Sigmoid transformation and Linear scaling $Pw$ to $[0, 1]'$.

But results did not improve.

The analytical solution of (3) is

$$w = (P^TP + \frac{\lambda}{2} I)^{-1} P^T y,$$

(5)

where $I$ is the identity matrix.

The problem is that $y$ is unknown.
Estimating $y$: First Approach

- Use validation data to estimate $w$.
- Training set $\Rightarrow V$ and $\tilde{V}$ internally
- Student sub-teams generated two prediction results on $\tilde{V}$ and $\tilde{T}$:
  
  \begin{align*}
  \text{Train } V & \Rightarrow \text{ Predict } \tilde{V} \text{ to obtain } \tilde{p}_i, \\
  \text{Train } T & \Rightarrow \text{ Predict } \tilde{T} \text{ to obtain } p_i.
  \end{align*}

- Let $\tilde{P}$ the matrix collecting all $\tilde{p}_i$; we know true $\tilde{y}$.
- In (3) using $\tilde{y}$ and $\tilde{P}$ to obtain $w$.
- Final prediction: we calculated $Pw$ and applied the truncation in (4).
Ensemble and Final Results

Estimating $y$: Second Approach

- Use leaderboard information to estimate $P^Ty$ in (5). We follow from Töscher and Jahrer (2009).

$$r_i \equiv \sqrt{\frac{\|p_i - y\|^2}{l}},$$

so

$$p_i^T y = \frac{\|p_i\|^2 + \|y\|^2 - lr_i^2}{2}. \quad (6)$$

- $r_i$ and $\|y\|$ unavailable; estimated by

$$r_i \approx \hat{r}_i \quad \text{and} \quad \|y\|^2 \approx l\hat{r}_0^2,$$

$\hat{r}_i$: RMSE on the leaderboard by submitting $p_i$

$\hat{r}_0$: RMSE by submitting the zero vector.
Ensemble Results

- We collect 19 results from 7 sub-teams
- Each result comes from training a single classifier
- To select $\lambda$, we gradually increased $\lambda$ until the leaderboard result started to decline
- This procedure, conducted in the last several hours before the deadline, was not very systematic
Best A89 result: $P^T y$ in (5) and using $\lambda = 10$. That is, second approach.

Best B89 result: using the validation set to estimate $w$ and $\lambda = 0$ (no regularization).

This means the first approach.
Ensemble Results (Cont’d)

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
Many submissions in the last week before the deadline; in particular in the last two hours

- Everyone (including ourselves) tries to achieve better leader board results
- Overfitting may be a concern
  Not very clear how serious this problem is
# Ensemble and Final Results

## Leaderboard Immediately After the Deadline

### Number of rows: 3419

<table>
<thead>
<tr>
<th>Overall Rank</th>
<th>Individual/Team Name</th>
<th>Algebra I 2008-2009</th>
<th>Bridge to Algebra 2008-2009</th>
<th>Total Score</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277687</td>
<td>0.27679</td>
<td>2010-06-06 11:42:46</td>
</tr>
<tr>
<td>2</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277687</td>
<td>0.27679</td>
<td>2010-06-06 11:44:06</td>
</tr>
<tr>
<td>3</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277691</td>
<td>0.276792</td>
<td>2010-06-08 23:22:22</td>
</tr>
<tr>
<td>4</td>
<td>BbCc</td>
<td>0.275898</td>
<td>0.277687</td>
<td>0.276793</td>
<td>2010-06-06 11:44:14</td>
</tr>
<tr>
<td>5</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277694</td>
<td>0.276793</td>
<td>2010-06-08 22:48:45</td>
</tr>
<tr>
<td>6</td>
<td>BbCc</td>
<td>0.275908</td>
<td>0.277687</td>
<td>0.276797</td>
<td>2010-06-08 23:37:58</td>
</tr>
<tr>
<td>7</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277703</td>
<td>0.276798</td>
<td>2010-06-08 23:30:14</td>
</tr>
<tr>
<td>8</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277707</td>
<td>0.2768</td>
<td>2010-06-08 23:03:36</td>
</tr>
<tr>
<td>9</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277712</td>
<td>0.276802</td>
<td>2010-06-08 23:18:58</td>
</tr>
<tr>
<td>10</td>
<td>National Taiwan University</td>
<td>0.275615</td>
<td>0.277991</td>
<td>0.276803</td>
<td>2010-06-08 23:46:50</td>
</tr>
<tr>
<td>11</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277713</td>
<td>0.276803</td>
<td>2010-06-08 22:54:56</td>
</tr>
<tr>
<td>12</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277718</td>
<td>0.276805</td>
<td>2010-06-08 22:59:44</td>
</tr>
<tr>
<td>13</td>
<td>National Taiwan University</td>
<td>0.275615</td>
<td>0.277998</td>
<td>0.276806</td>
<td>2010-06-08 23:34:02</td>
</tr>
<tr>
<td>14</td>
<td>National Taiwan University</td>
<td>0.275615</td>
<td>0.277998</td>
<td>0.276806</td>
<td>2010-06-08 23:37:55</td>
</tr>
<tr>
<td>15</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.27772</td>
<td>0.276807</td>
<td>2010-06-08 23:11:23</td>
</tr>
<tr>
<td>16</td>
<td>BbCc</td>
<td>0.275927</td>
<td>0.277867</td>
<td>0.276807</td>
<td>2010-06-08 23:32:15</td>
</tr>
<tr>
<td>17</td>
<td>National Taiwan University</td>
<td>0.275617</td>
<td>0.277998</td>
<td>0.276807</td>
<td>2010-06-08 22:57:55</td>
</tr>
<tr>
<td>18</td>
<td>BbCc</td>
<td>0.275893</td>
<td>0.277728</td>
<td>0.276811</td>
<td>2010-06-08 23:16:02</td>
</tr>
<tr>
<td>19</td>
<td>BbCc</td>
<td>0.275935</td>
<td>0.277694</td>
<td>0.276815</td>
<td>2010-06-08 20:55:26</td>
</tr>
<tr>
<td>20</td>
<td>BbCc</td>
<td>0.275935</td>
<td>0.277694</td>
<td>0.276815</td>
<td>2010-06-08 22:45:33</td>
</tr>
</tbody>
</table>
**Final submissions of all teams with a fact sheet**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Cup Score</th>
<th>Leaderboard Score</th>
<th>Final Submission Time</th>
<th>Fact Sheet</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National Taiwan University</td>
<td>0.272952</td>
<td>0.276803</td>
<td>2010-06-08 23:46:50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Zhang and Su</td>
<td>0.273692</td>
<td>0.276790</td>
<td>2010-06-08 23:39:35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BigChaos @ KDD</td>
<td>0.274556</td>
<td>0.279046</td>
<td>2010-06-07 03:48:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Zach A. Pardos</td>
<td>0.276590</td>
<td>0.279695</td>
<td>2010-06-08 21:31:07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Old Dogs With New Tricks</td>
<td>0.277864</td>
<td>0.281163</td>
<td>2010-06-08 23:49:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SCUT Data Mining</td>
<td>0.280476</td>
<td>0.284624</td>
<td>2010-06-08 23:25:27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>pinta</td>
<td>0.284550</td>
<td>0.289200</td>
<td>2010-06-08 22:14:55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Course at NTU
- 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
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

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.