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

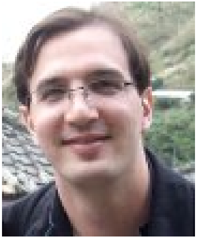
Team Members

- 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)

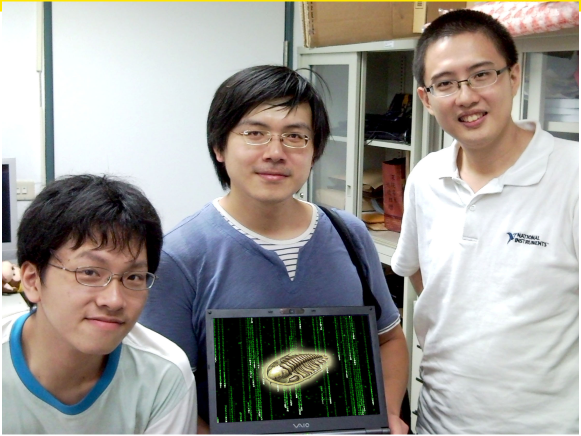
Starfish

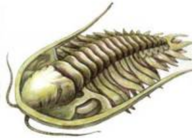
Chia-Hua Ho (何家華), Po-Han Chung (鐘博翰), and Jung-Wei Chou (周融瑋)

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)

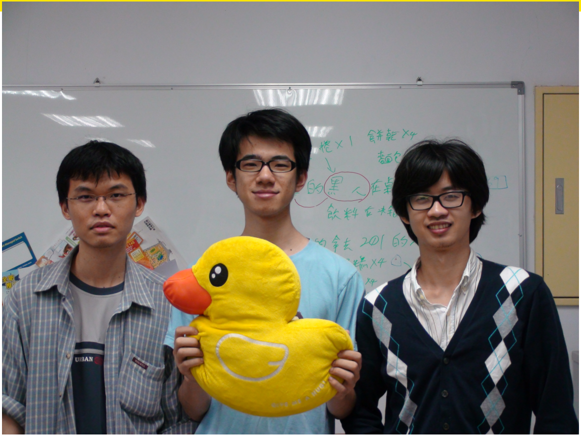
Trilobite





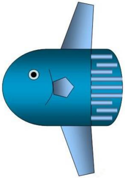
Yi-Chen Lo (羅亦辰), Che-Wei Chang (張哲維) and Tsung-Ting Kuo (郭宗廷)

Duck



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

Sunfish

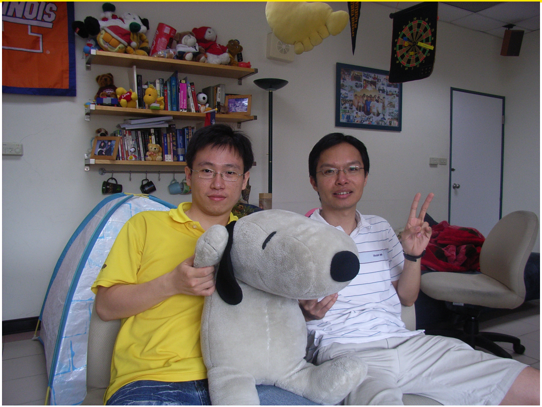


Yu-Xun Ruan (阮昱勳), Chen-Wei Hung (洪琛洧) and Yi-Hung Huang (黃曳弘)

Tiger (RA)

Yu-Shi Lin (林育仕)

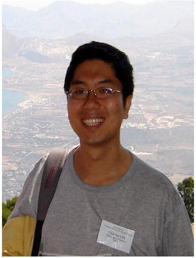
Snoopy (TAs)

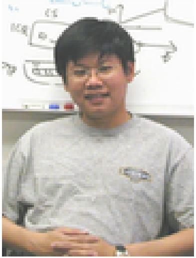
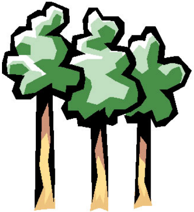


Hsiang-Fu Yu (余相甫) and Hung-Yi Lo (駱宏毅)

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

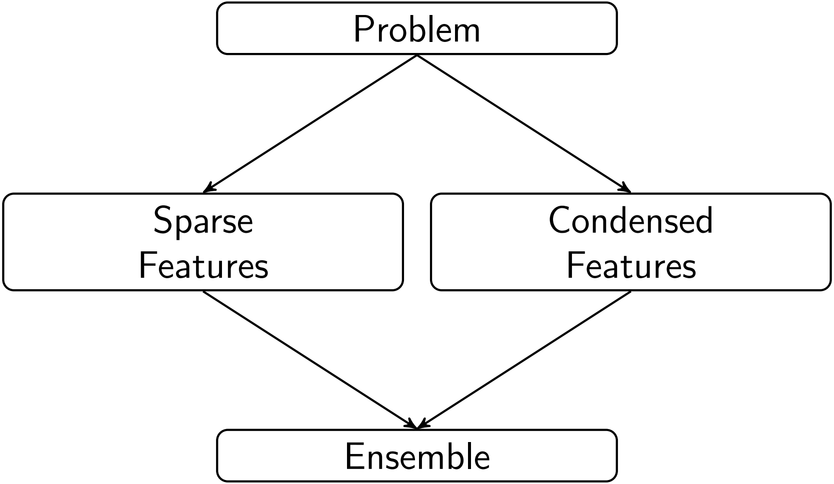
Initial Thoughts and Our Approach (Cont'd)

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



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

A unit of problems

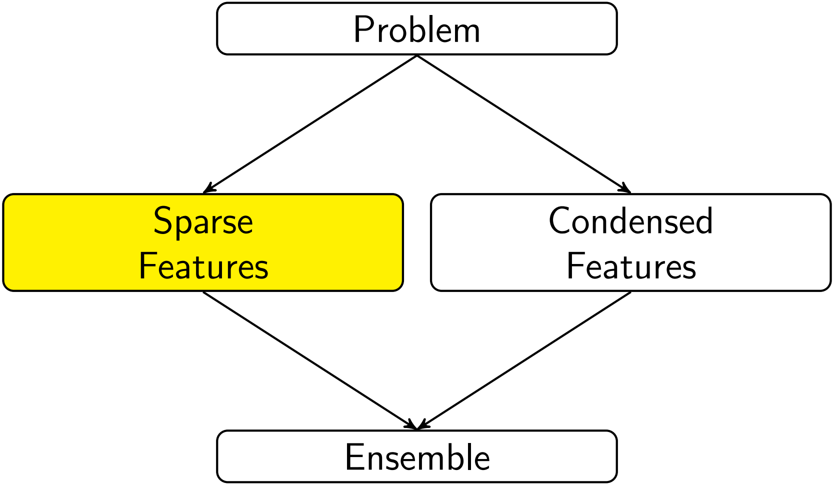
problem 1 $\in V$
problem 2 $\in V$
$\vdots$
last problem $\in \tilde{V}$

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


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

RMSE (leader board)	A89	B89
Basic sparse features	0.2895	0.2985
Best leader board	0.2759	0.2777

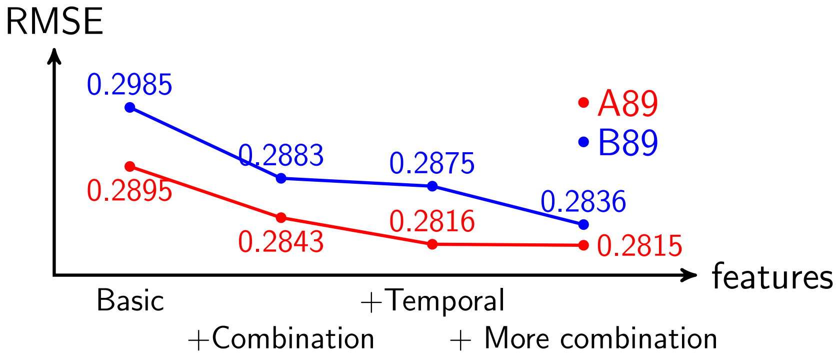
Five of six student sub-teams use variants of this approache

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)



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

	#instances	#features
A89	8,918,055	$\geq$ 20M
B89	20,012,499	$\geq$ 30M

- 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)

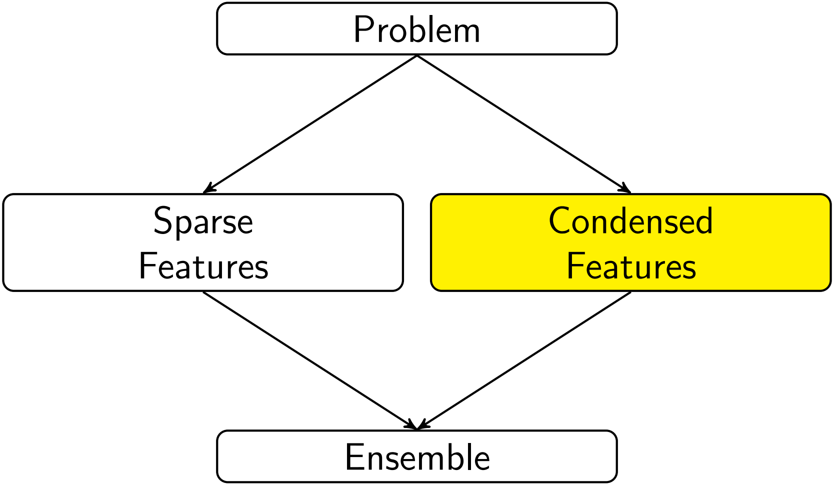
Result Using Sparse Features

Leader board results:

	A89	B89
Basic sparse features	0.2895	0.2985
Best sparse features	0.2784	0.2830
Best leader board	0.2759	0.2777

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

Condensed Features

Categorical feature ⇒ numerical feature

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

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

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

Temporal feaures: the previous ≤ 6 steps with the same student and KC

- An indicator for the existence of such steps
- Correct first attempt rate
- Average hint request rate

Condensed Features (Cont'd)

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

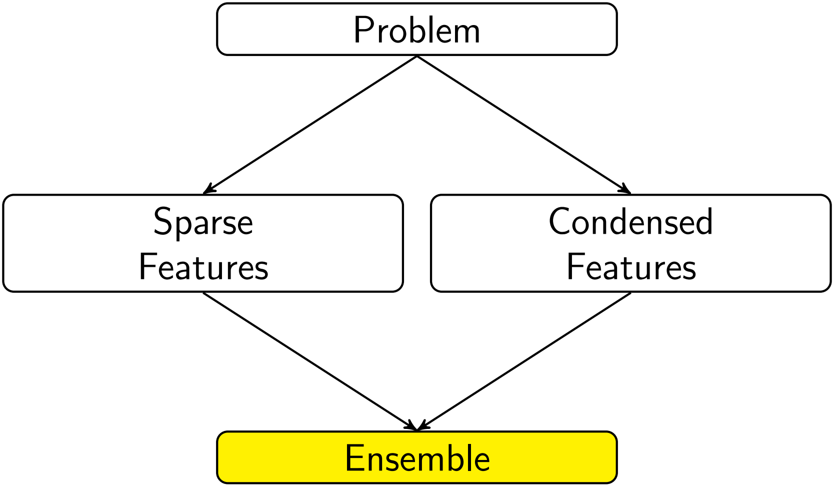
Training by Random Forest

- 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:

	A89	B89
Basic sparse features	0.2895	0.2985
Best sparse features	0.2784	0.2830
Best condensed features	0.2824	0.2847
Best leader board	0.2759	0.2777

- This small feature set works well

Outline


Problem
Sparse Features
Condensed Features
Ensemble

Linear Regression for Ensemble

Linear regression to ensemble sub-team results

$$\min_{\mathbf{w}} \quad \|\mathbf{y} - P\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

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

Decision of the regularization parameter λ

Ensemble Results

Ensemble significantly improves the results

	A89	B89	Avg.
Basic sparse features	0.2895	0.2985	0.2940
Best sparse features	0.2784	0.2830	0.2807
Best condensed features	0.2824	0.2847	0.2835
Best ensemble	0.2756	0.2780	0.2768
Best leader board	0.2759	0.2777	0.2768

- 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

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

- Team names used during the competition:
  Snoopy ⇒ National Taiwan University
  BbCc ⇒ 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