



Learning Rules for Prioritizing Investigations of Public Companies Suspected of Financial Fraud

Jianping Zhang Ali Hadjarian Brent Han Jerzy Bala
The MITRE Corporation
August 24, 2008

Outline

- **Financial Fraud Challenge**
- **Requirements and Constraints**
- **Rule Induction**
- **Ranking with Rules**
- **Performance Evaluation**
- **Experimental Results**
- **Conclusions and Future Work**

Financial Fraud Challenge

Problem

Investigation of fraudulent activities is a challenging task due to limited availability of resources required for in-depth investigations

- *Despite increasingly stringent legislation aimed at combating fraud, **financial fraud** remains a public concern*
 - ❑ Internal financial fraud increased by 19% in 2006 (Global Security Survey, Deloitte)
 - ❑ U.S. organizations will lose \$994 billion in 2008 due to fraud (2008 Report to the Nation, ACFE)



Research Objective

Development of predictive models for prioritizing human investigations of companies suspected of engaging in fraudulent behavior based on their financial filings

Requirements & Constraints

- **Optimizing the use of limited inspection resources**
 - Need predictive models with ranking capabilities
 - Emphasis should be on highest ranked targets
- **Transparency**
 - Models should be human comprehensible and provide insight to inspectors
- **Imbalanced class distributions**
 - Only a small portion of the data deals with fraudulent cases
- **Unlabeled training cases**
 - Most cases have not been labeled by human inspectors
- **Changing patterns**
 - Patterns of suspicious activities could change over time

Rule Induction Algorithm

- Learns rules for the minority class only
- General-to-specific heuristic beam search
- Dynamic discretization of numeric attributes
- Uses F-measure as the metric for rule evaluation

$$F - measure(r) = \frac{\beta^2 + 1}{\frac{\beta^2}{recall(r)} + \frac{1}{precision(r)}}$$

- Parameters
 - β : favor recall when $\beta > 1$ and favor precision when $\beta < 1$
 - Minimum Recall
 - Minimum Precision

Ranking with Rules

- Rules typically allow for binary decisions only
- Probability estimation can be used for score assignment

$$\frac{k+1}{n+C}$$

where k and n are the number of positive examples and the total number of examples covered, respectively, and C is the number of classes.

- Computing the score of an example covered by multiple rules

- Max:

$$score(e, RS) = \max_{i=1}^l \{score(e, R_i)\}$$

- Average

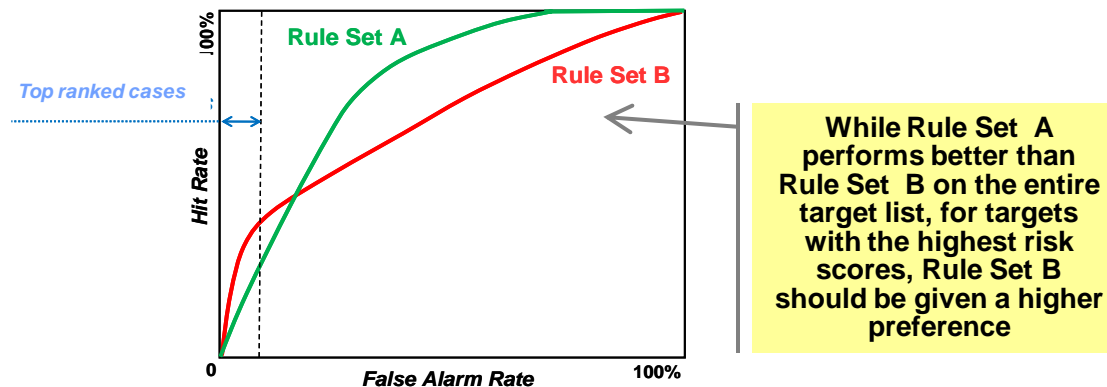
$$score(e, RS) = \frac{\sum_{i=1}^l score(e, R_i)}{l}$$

- Probabilistic Sum

$$score(e, \{R_1, R_2\}) = score(e, R_1) + score(e, R_2) - score(e, R_1) \times score(e, R_2)$$

Performance Evaluation

- Due to limited resources, only a tiny percentage of top ranked suspicious companies (1% to 2%) may be selected for further inspections; Preference should be given to models that **optimize the performance on the top ranked cases** as opposed to the entire dataset
- The performance of a risk scoring model on the entire target list is not necessarily consistent with its performance on the targets with the highest risk scores (as captured by the area under the leftmost part of the ROC curve or LAUC)



Experimental Dataset

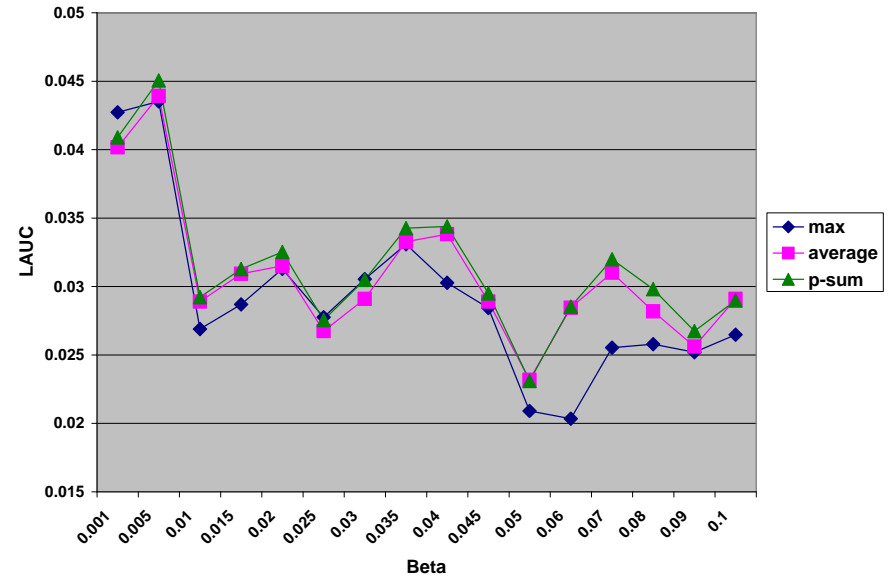
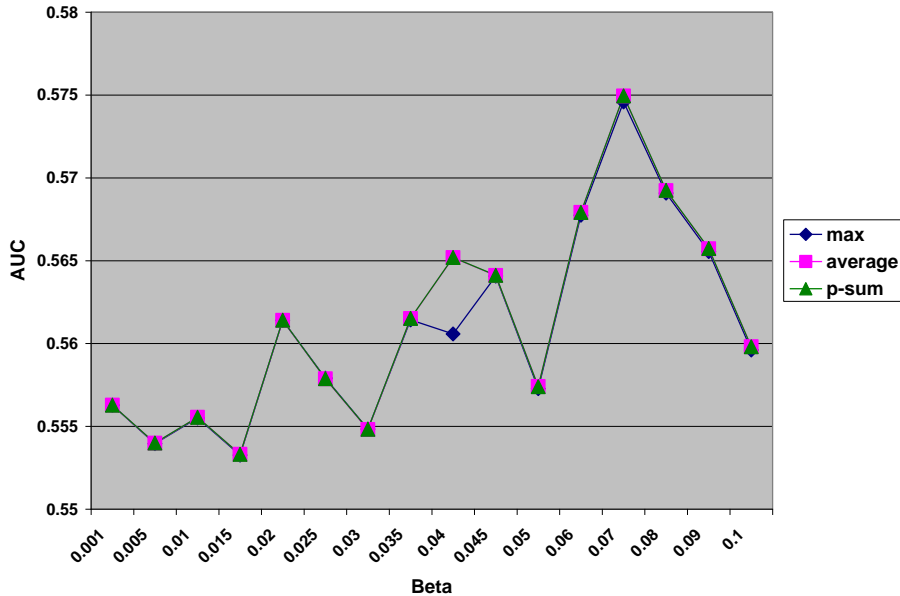
- Data Source
 - Publicly available SEC EDGAR filings
- Attributes
 - Around 130 attributes (95 numeric and 35 symbolic): a combination of original data as well as calculated and derived attributes
- Examples
 - Positive examples are filings for companies that had issued material restatements
 - Negative examples are filings for companies that had not issued restatements
 - Training data: FY 2003 – FY 2004; 4,932 positive and 38,792 negative examples
 - Test data: FY 2005; 836 positive and 18,780 negative examples

Experimental Setting

- Validation
 - 5 different runs of each experiment, using random sampling of the data
- Rule Learning Parameter Settings
 - Varying values for β
 - Varying values for minimum recall and minimum precision
- Rule-based Scoring
 - Probability estimation
 - Score combining methods
 - ❖ Max
 - ❖ Average
 - ❖ Probabilistic Sum
- Evaluation Metrics
 - AUC
 - LAUC at 1% false positive rate cutoff point

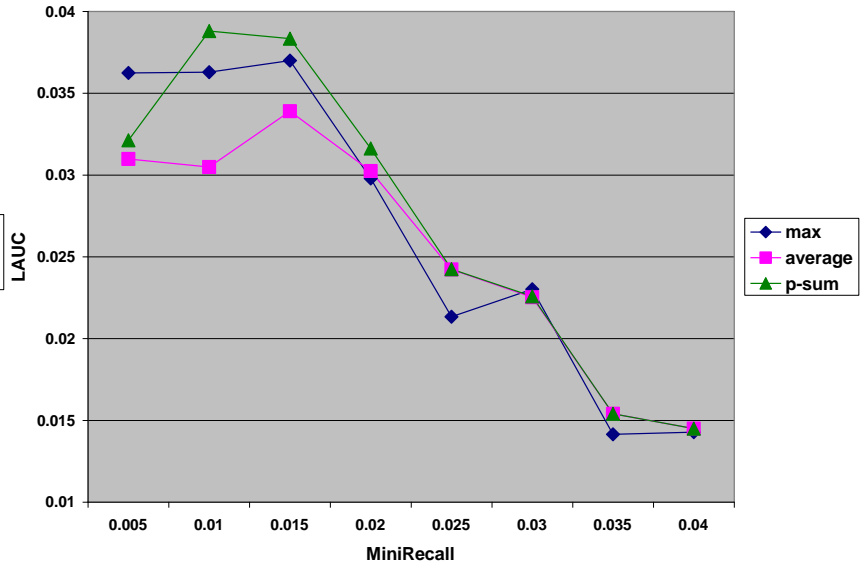
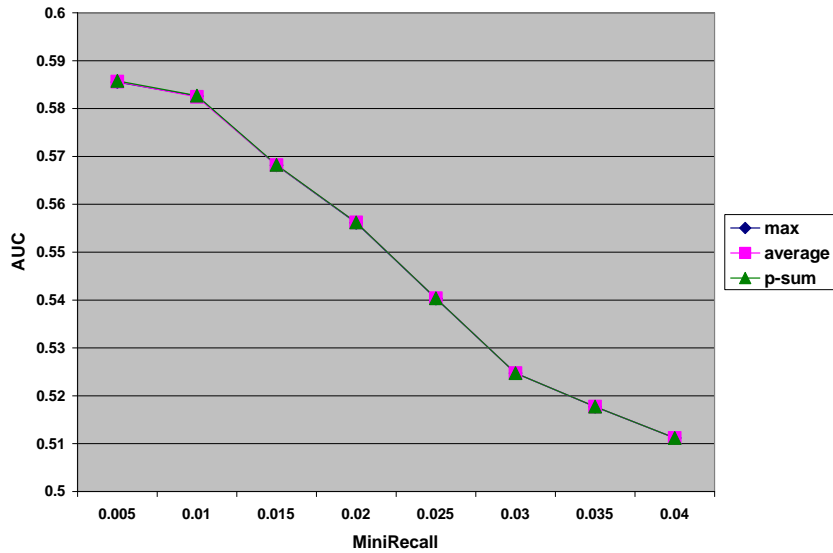
Experimental Results: I

- β : from 0.001 to 0.1
- Minimum recall = 0.01 and minimum precision = 0.6



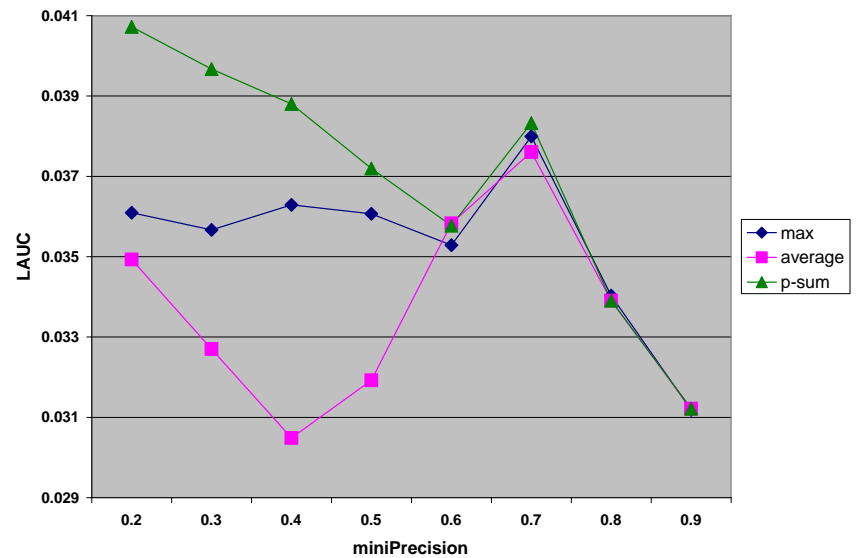
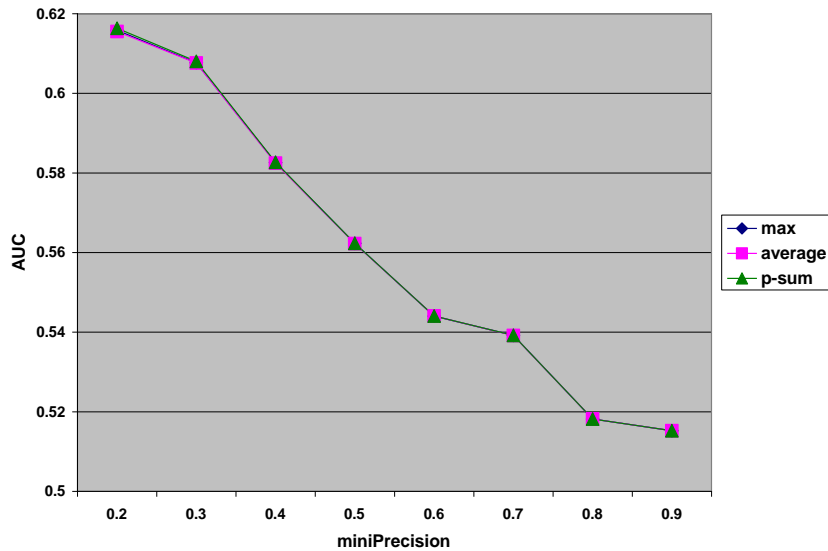
Experimental Results: II

- Minimum recall: from 0.005 to 0.04
- minimum precision = 0.4 and $\beta = 0.001$



Experimental Results: III

- Minimum precision: from 0.2 to 0.9
- minimum recall = 0.01 and $\beta = 0.001$



Conclusions and Future Work

- Conclusions
 - LAUC does not necessarily correlate with AUC
 - Various inductive biases could impact LAUC and AUC differently
 - ❖ E.g., Specific and precise rules result in higher LAUC values, while more general rules result in higher AUC values
 - The three score combining methods yielded about the same AUC values, but *P-sum* generally resulted in higher LAUC values
- Future Work
 - Develop other rule-based scoring methods
 - Develop a rule induction algorithm for optimizing the performance on top ranked cases