Improving Cause Detection Systems with Active Learning

Isaac Persing and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Cause Identification

- determines **why** the incident described in an incident report in the ASRS database occurred

- A **text categorization** task
  - NASA researchers have identified 14 causes (or *shaping factors*) that could explain why an incident occurred
  - **Goal**: given an incident report, determine which of a set of 14 shapers contributed to the occurrence of the incident

Shaping Factors (Posse et al., 2005)

- **Proficiency**
  - general deficit in capabilities
    - inexperience, lack of training, not qualified, …
- **Physical Factors**
  - pilot ailment that could impair flying
    - being tired, drugged, ill, dizzy, …
- **Resource Deficiency**
  - absence, insufficient number, or poor quality of a resource
    - overworked or unavailable controller, insufficient or out-of-date chart, malfunctioning or missing equipment

Shaping Factors (Cont')

- **Attitude**
- **Physical Environment**
- **Communication Environment**
- **Familiarity**
- **Pressure**
- **Preoccupation**
- **Taskload**
- **Duty Cycle**
- **Illusion**
- **Unexpected**
- **Other**

Cause Identification is Challenging

- **No publicly available labeled data**
- **Skewed class distributions**
  - some shapers occur a lot more frequently than the others
  - 10 of the 14 shapers are minority classes
- **Multi-label categorization**
  - an incident may be caused by more than one factor

Cause Identification is Challenging

- **No publicly available labeled data**

- **Skewed class distributions**
  - some shapers occur a lot more frequently than the others
  - 10 of the 14 shapers are minority classes

- **Multi-label categorization**
  - Recast the 14-class classification task as a set of 14 binary tasks
  - Train each binary (SVM) classifier using a one-vs-all scheme
  - Each report may receive one or more labels

Cause Identification is Challenging

- **No publicly available labeled data**

- **Skewed class distributions**
  - Reduce data skewness by oversampling

Cause Identification is Challenging

- **No publicly available labeled data**

Goal: Improve cause identification by reducing the cost of data annotation via active learning

Dataset (1,333 Hand-Labeled Reports)

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Dataset (1,333 Hand-Labeled Reports)

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Dataset (1,333 Hand-Labeled Reports)

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Minority Shapers

Goal

- Improve cause identification by reducing data annotation cost via active learning

Active Learning

- Have a human annotator annotate only those unlabeled instances that are most informative to the machine learner
  - **Most informative instances**
    - instances whose label the learner is most uncertain about
  - **Margin-based active learning**
    - use an SVM learner to learn a hyperplane
    - unlabeled instances closest to hyperplane are most informative

Margin-Based Active Learning

Input: U: a large pool of unlabeled reports

1. Select 14 reports randomly from U and hand annotate them

2. Train 14 binary SVM classifiers on these labeled reports
   - one classifier for each shaper, using the one-vs-all scheme
   - each report is represented as a vector of unigrams (0/1)

3. Repeat
   - for each hyperplane, select the unlabeled report closest to it
   - hand-label these 14 newly selected reports
   - retrain the 14 classifiers on all of the reports annotated so far

Goal

- Improve this **Margin** baseline by investigating **four** extensions to the active learning framework

Extension 1: Oversampling

- **Motivation**
  - Because each binary SVM classifier is trained using a one-versus-all scheme, the training set exhibits class skewness
    - Positive instances outnumbered by negative instances
- **Solution**
  - Reduce class skewness by creating synthetic positive instances, as in the BootOS system (Zhu & Hovy, 2007)
- Each binary SVM classifier is trained on an oversampled version of the labeled data set in each active learning iteration

Extension 2: Overall Most Confident

- **Motivation**
  - The Margin baseline selects one report **per classifier** on each iteration, but it may be better to select reports that would be beneficial to multiple binary SVM classifiers.
  - Relax the "one report per classifier" constraint in the baseline

- **Idea** behind Overall Most Confident (**OMC**)
  - Exploits the multi-labelness of the cause identification task
  - On each iteration, it selects the 14 unlabeled reports that *N* of the 14 SVM classifiers are least confident about
    - If N=1, we call this extension OMC-1
    - If N=2, we call this extension OMC-2
    - modify the way we assign confidence values to the reports

Extension 3: Explore All Words

- **Motivation**
  - A good (labeled) training set should contain all relevant features to the task being learned.
  - Given that we have a small amount of labeled data, it is unlikely that we can identify all the relevant features.
  - The Explore All Words (EAW) extension prefers unlabeled reports containing many unseen words.

Four different versions of EAW

- Version 1: EAW
  - select the 14 unlabeled reports that contain the largest number of unseen unigrams with respect to the set of labeled reports
- Version 2: EAW-df
  - same as Version 1, but weigh each unigram by its document frequency computed over the set of unlabeled reports
    - Unigrams that appear more frequency may be more important
- Version 3: EAW-tfidf
  - Same as Version 1, but weigh each unigram by its tf-idf value
- Version 4: EAW-tfidf-df
  - combines versions 2 and 3

Extension 4: Document Length

- Motivation
  - Length of a report may tell us something about how desirable it is to have a report labeled
  - But … we are unsure whether we should prefer long or short documents.
    - a short report is less expensive to annotate
    - a long report tends to be associated with more shaping factors
      - provide useful positive instances for multiple binary classifiers

Two versions of Document Length

- Short version
  - select the 14 shortest reports for labeling in each iteration
- Long version
  - select the 14 longest reports for labeling in each iteration

Combining the four extensions

- Extensions 2-4 do not have to be used in isolation.

- How to combine them?
  - Scale the values by each extension to the range of 0 to 1
  - Assign each unlabeled report an overall confidence value that is equal to the sum of the values given by these extensions
  - Select the reports with the lowest confidence values

Evaluation

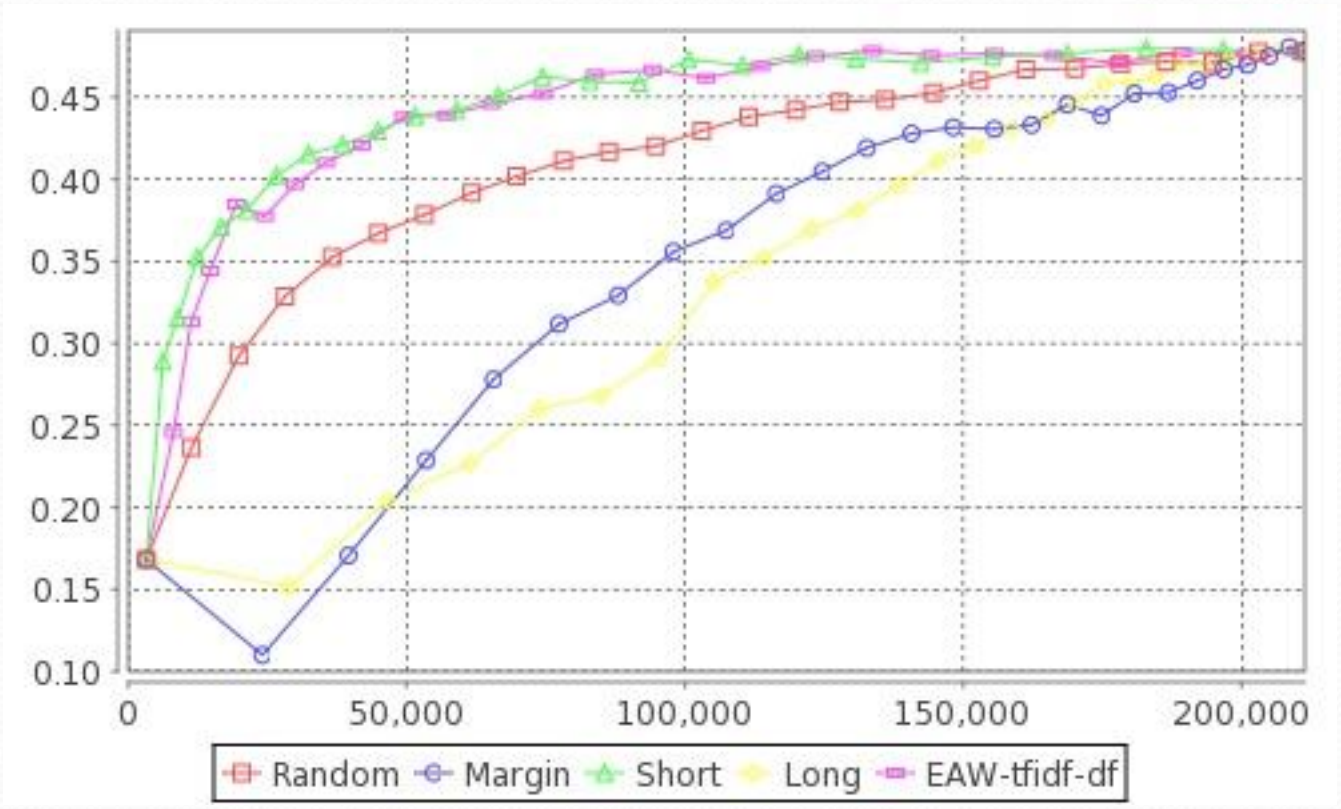
- 1,333 reports hand-labeled with shaper factors
- 5-fold cross validation
  - using one fold for testing
  - using as unlabeled data reports from the remaining four folds
- Results reported in the form of learning curves
  - F-measure scores micro-averaged over the 14 classes for different amounts of labeled data
- Two baselines
  - Margin
  - Random (passive learner)

Evaluation Goal

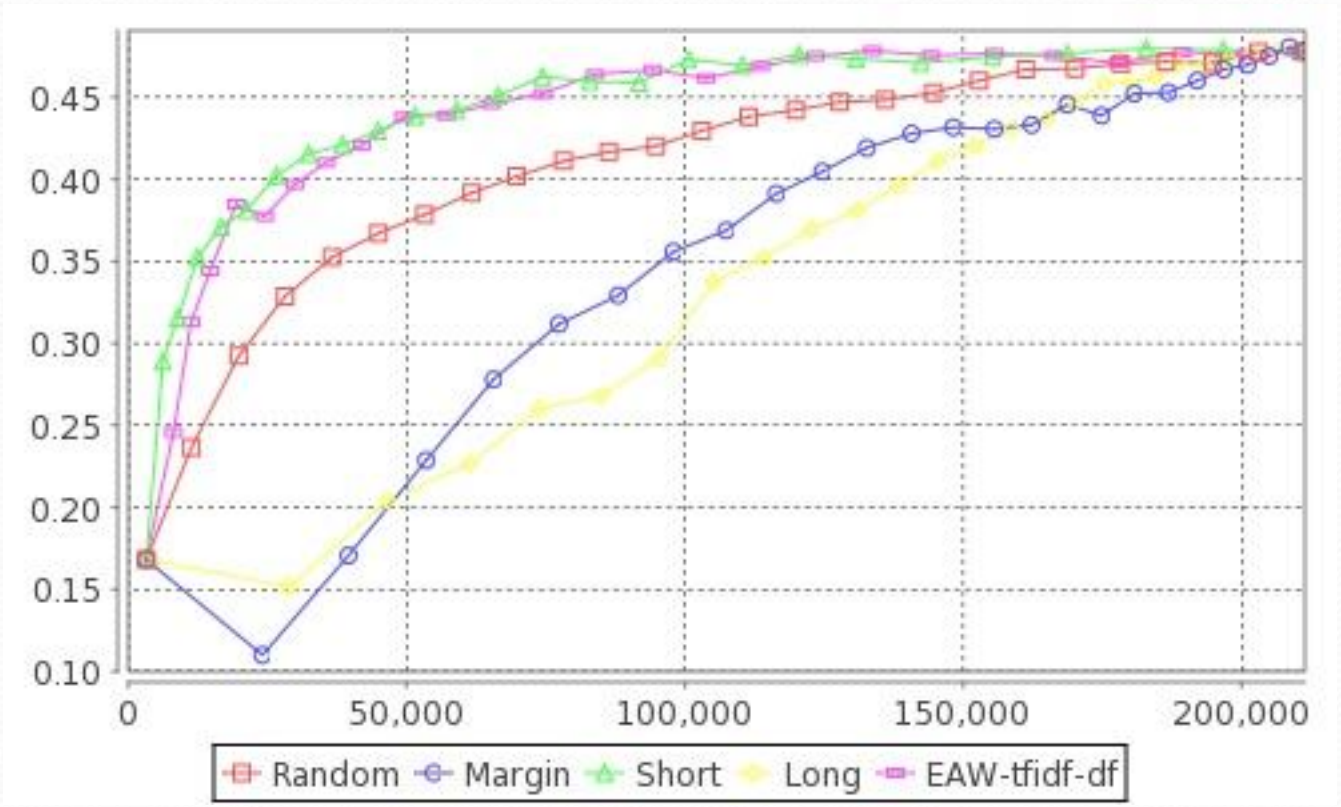
- measure the contribution of each extension to performance

- How?
  1. Start with an active learner that makes use of some version of all four extensions
     - Margin baseline + oversampling + OMC-1 + EAW-tfidf-df + Short
  2. Remove the extensions one at a time and observe the effects

Examining Extension 4 (Doc Length)

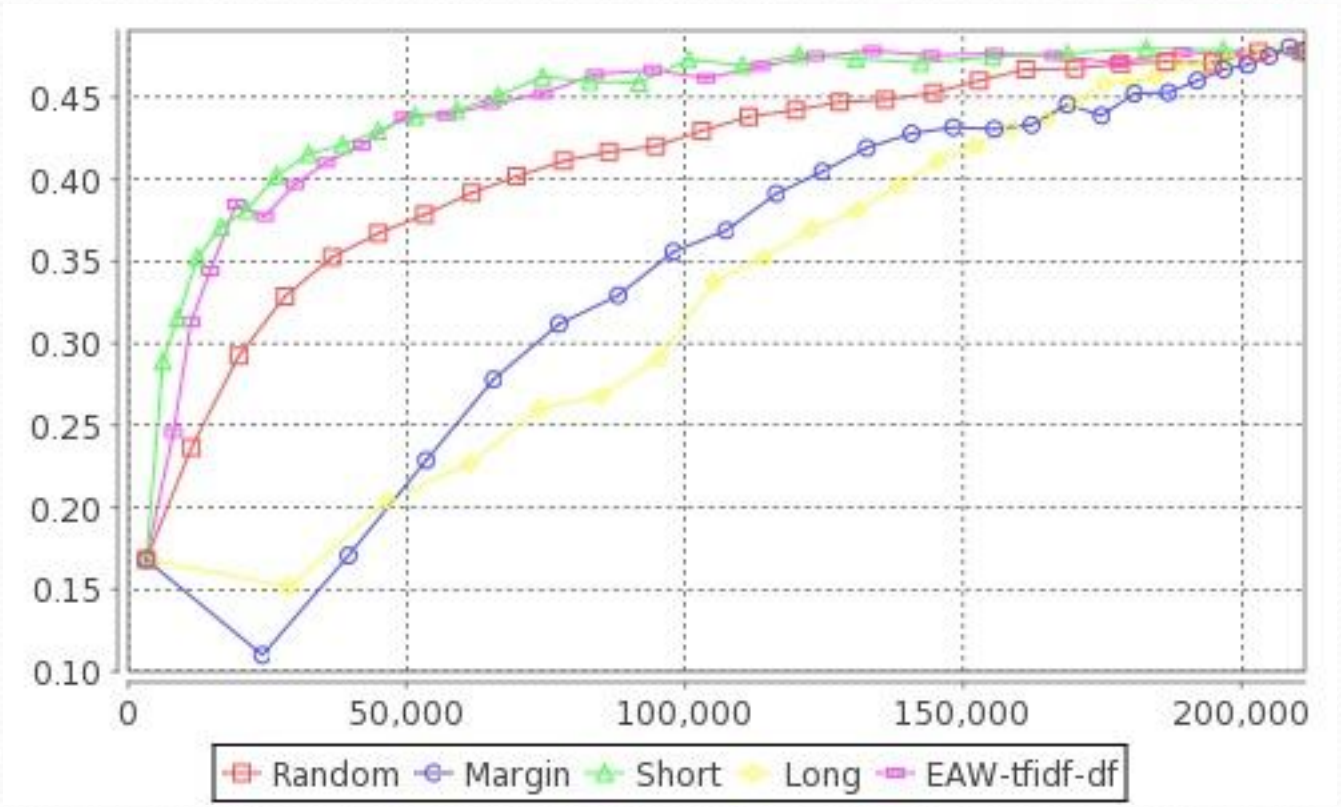


Examining Extension 4 (Doc Length)



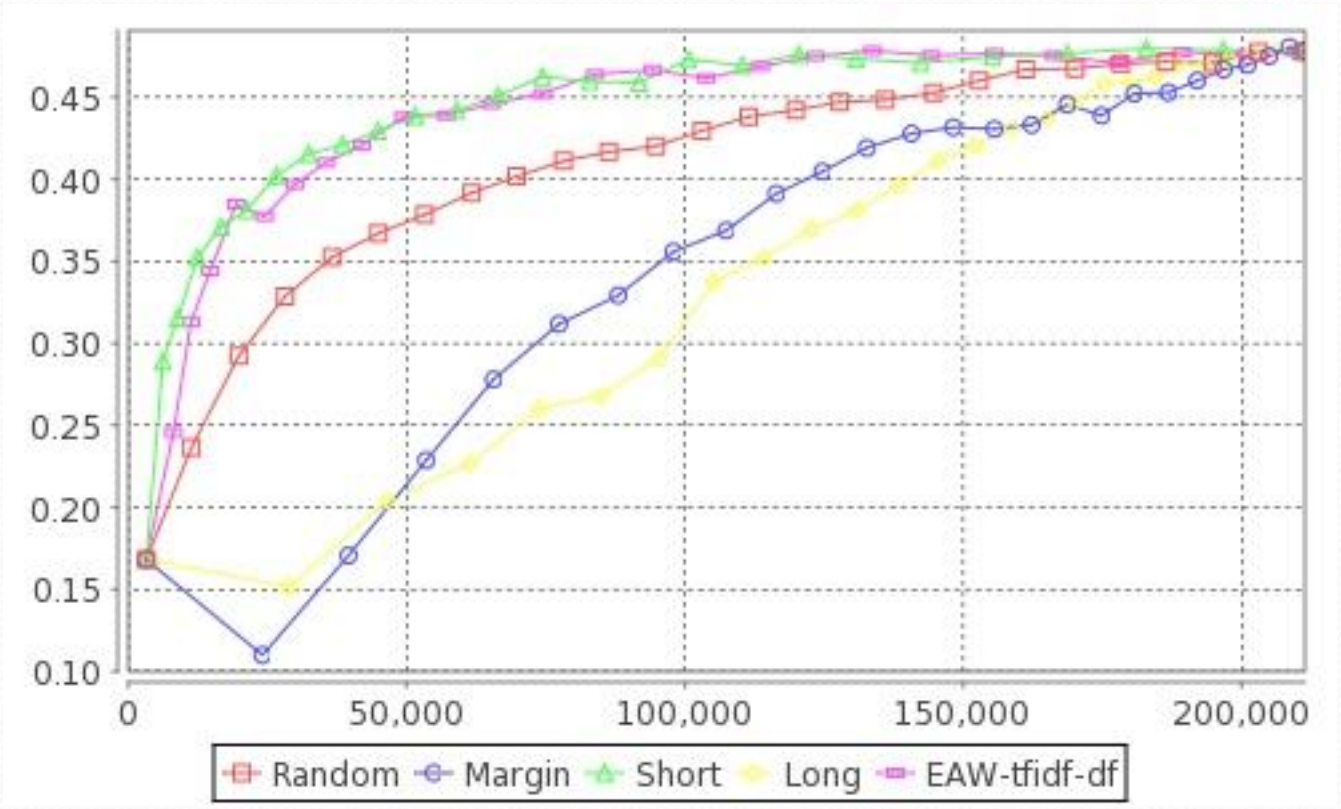
- y-axis: F-measure; x-axis: number of words in labeled reports

Examining Extension 4 (Doc Length)



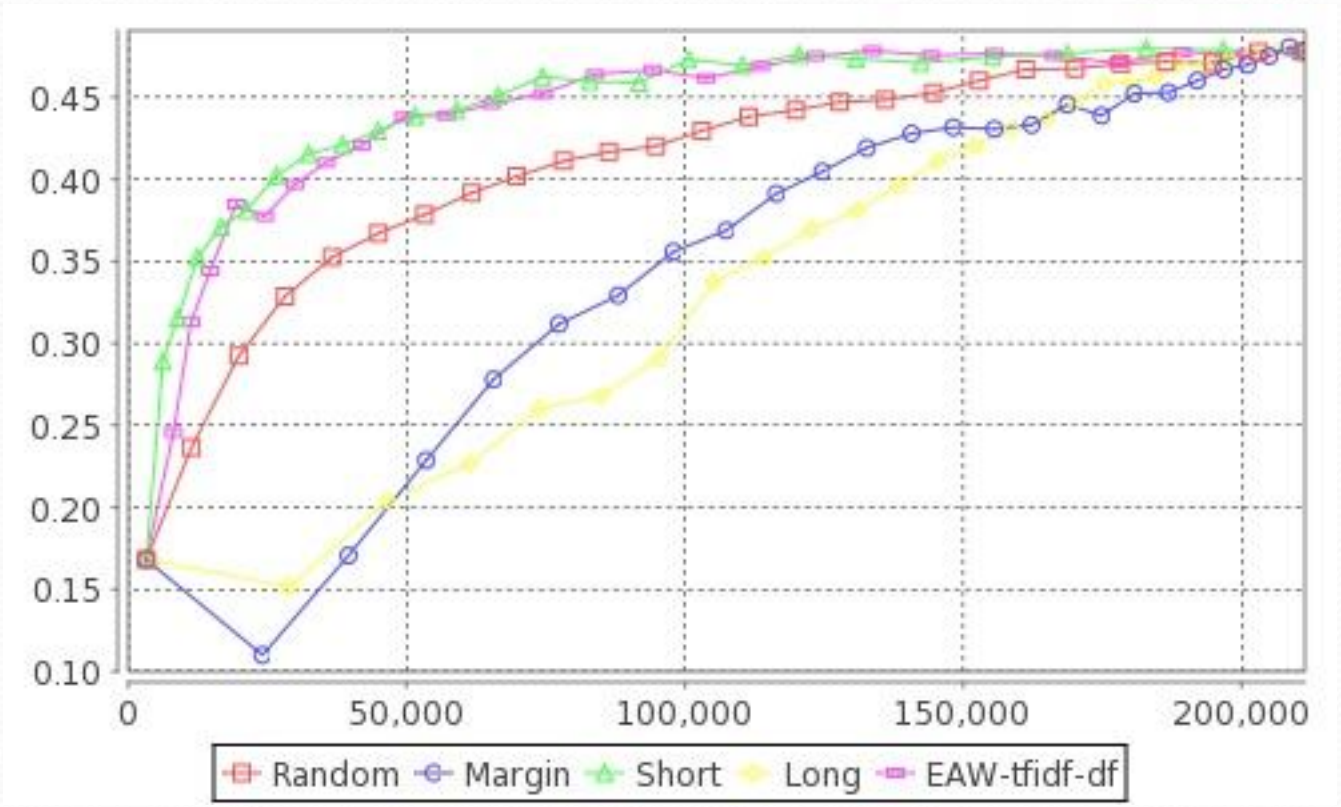
- Short (Green) and EAW-tfidf-df (Pink) perform the best

Examining Extension 4 (Doc Length)



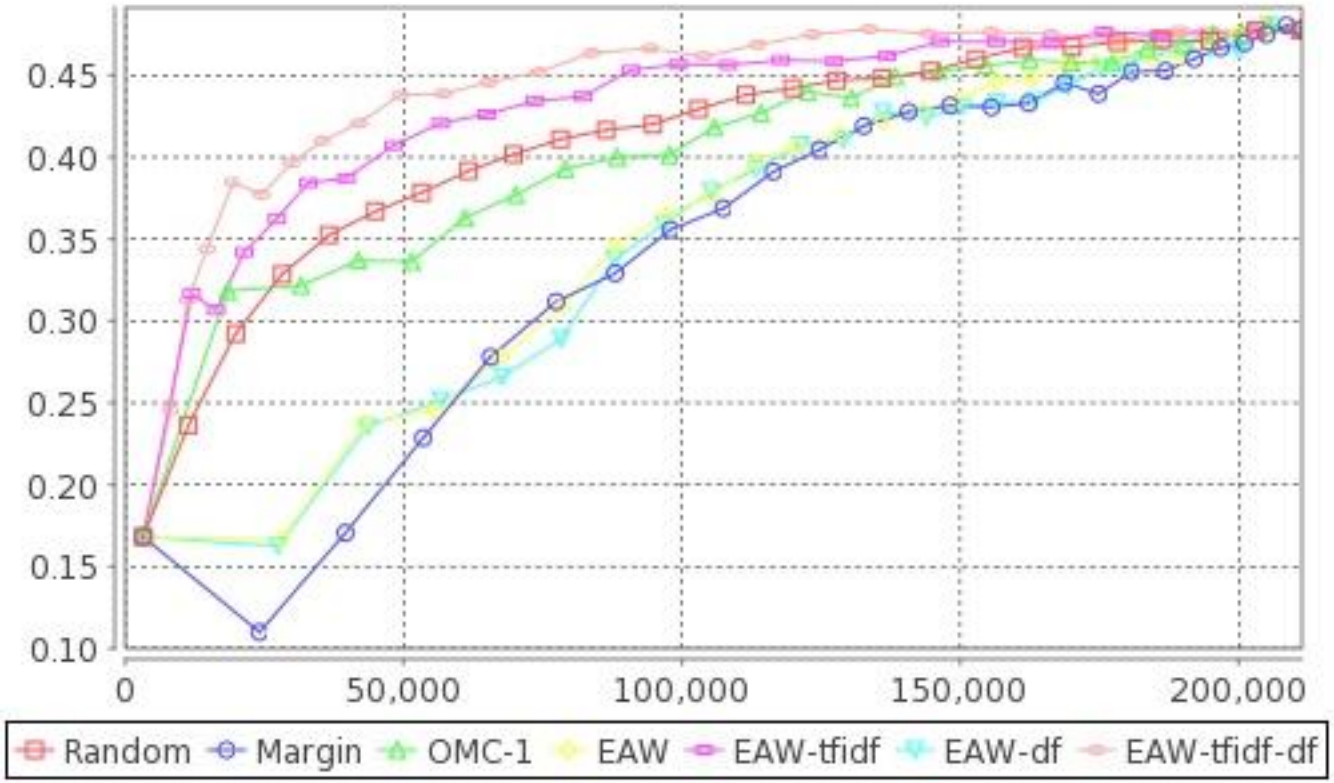
- Short (Green) and EAW-tfidf-df (Pink) perform the best
  - EAW-tfidf-df seems to have a built-in preference for short reports

Examining Extension 4 (Doc Length)



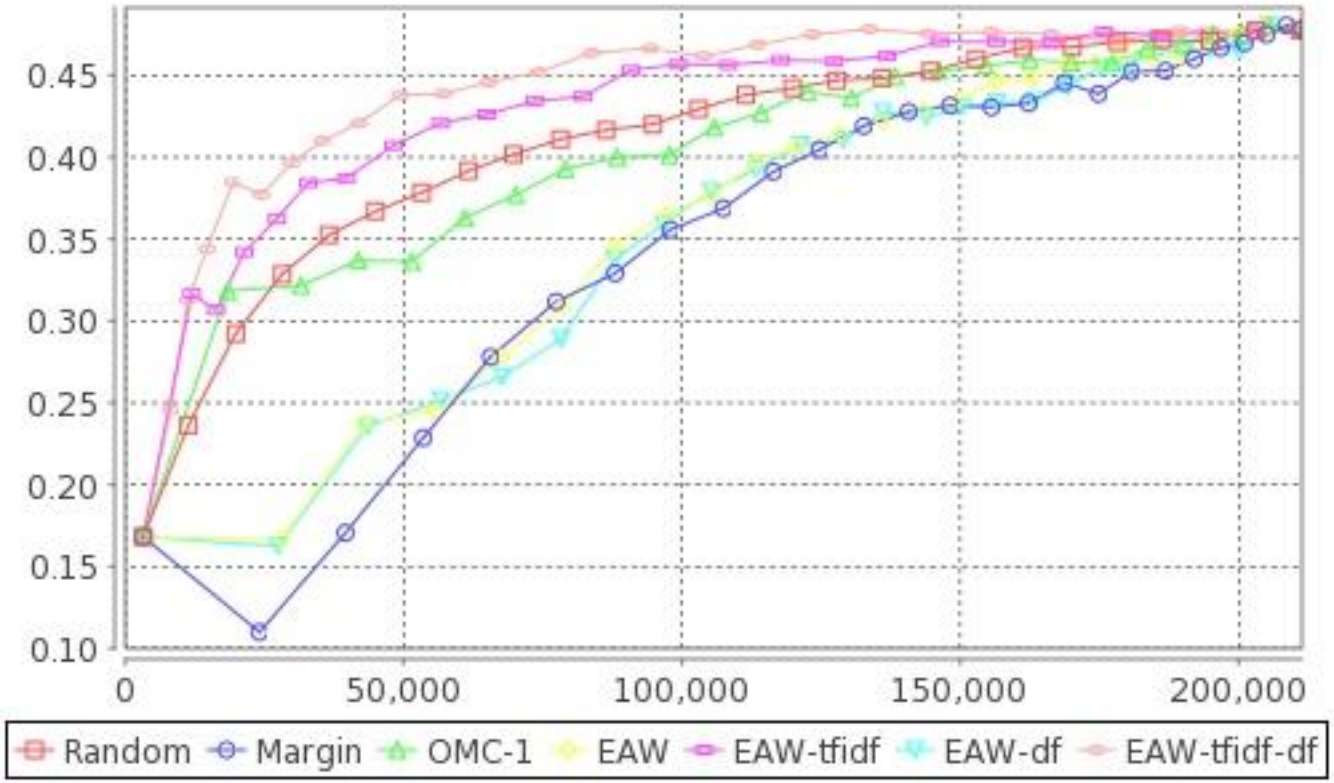
- Long (Yellow) is the worst
  - Long reports may contain info irrelevant to cause identification

Examining Extension 3 (EAW)



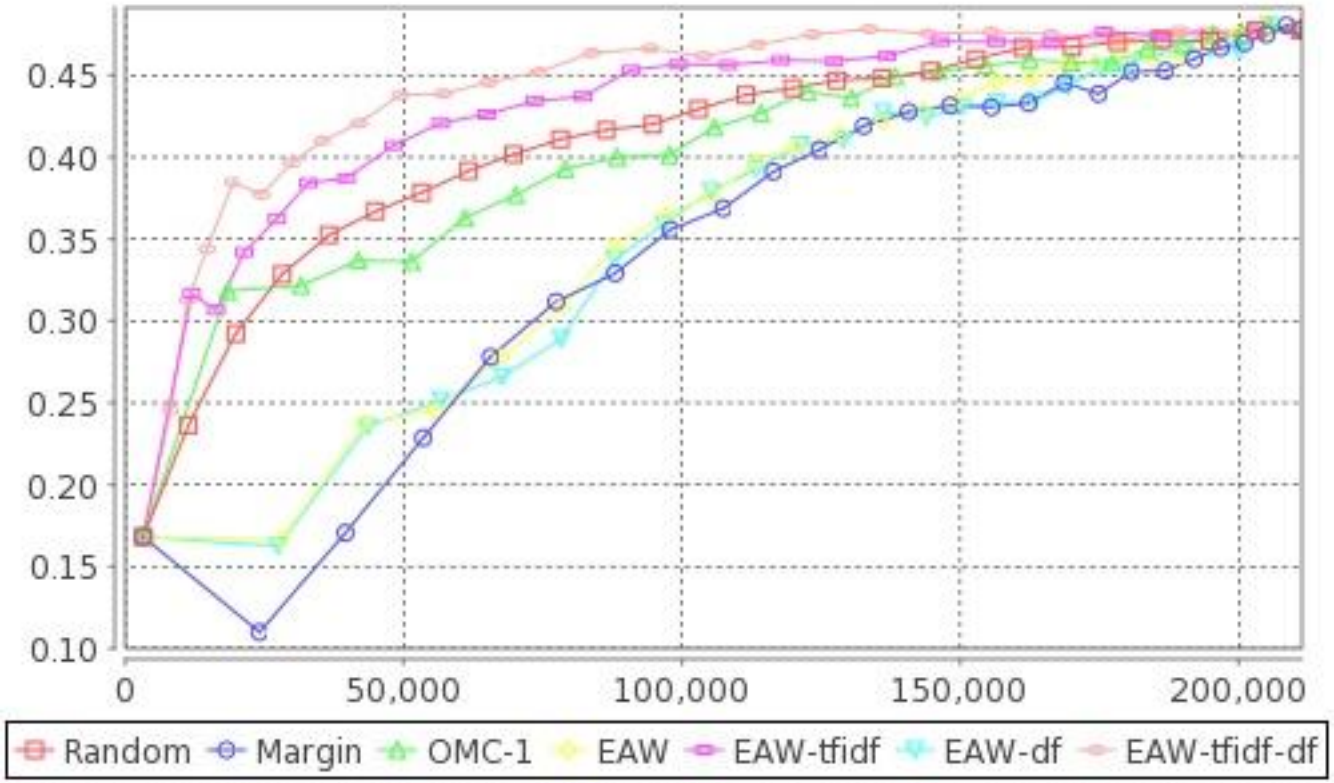
- The EAW extension prefers reports with many words not seen in the labeled set

Examining Extension 3 (EAW)



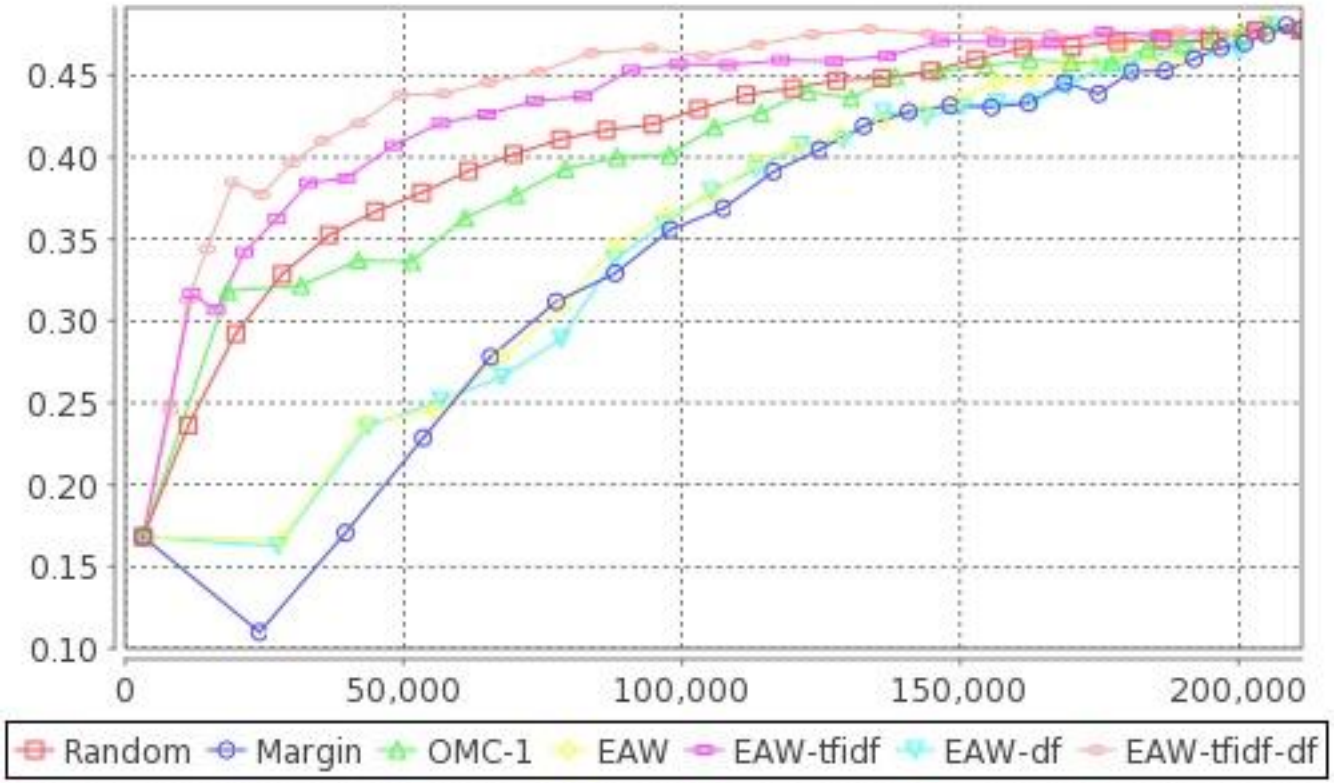
- 4 versions → 4 ways of assigning weights to unseen words
  - EAW, EAW-df, EAW-tfidf, EAW-tfidf-df

Examining Extension 3 (EAW)



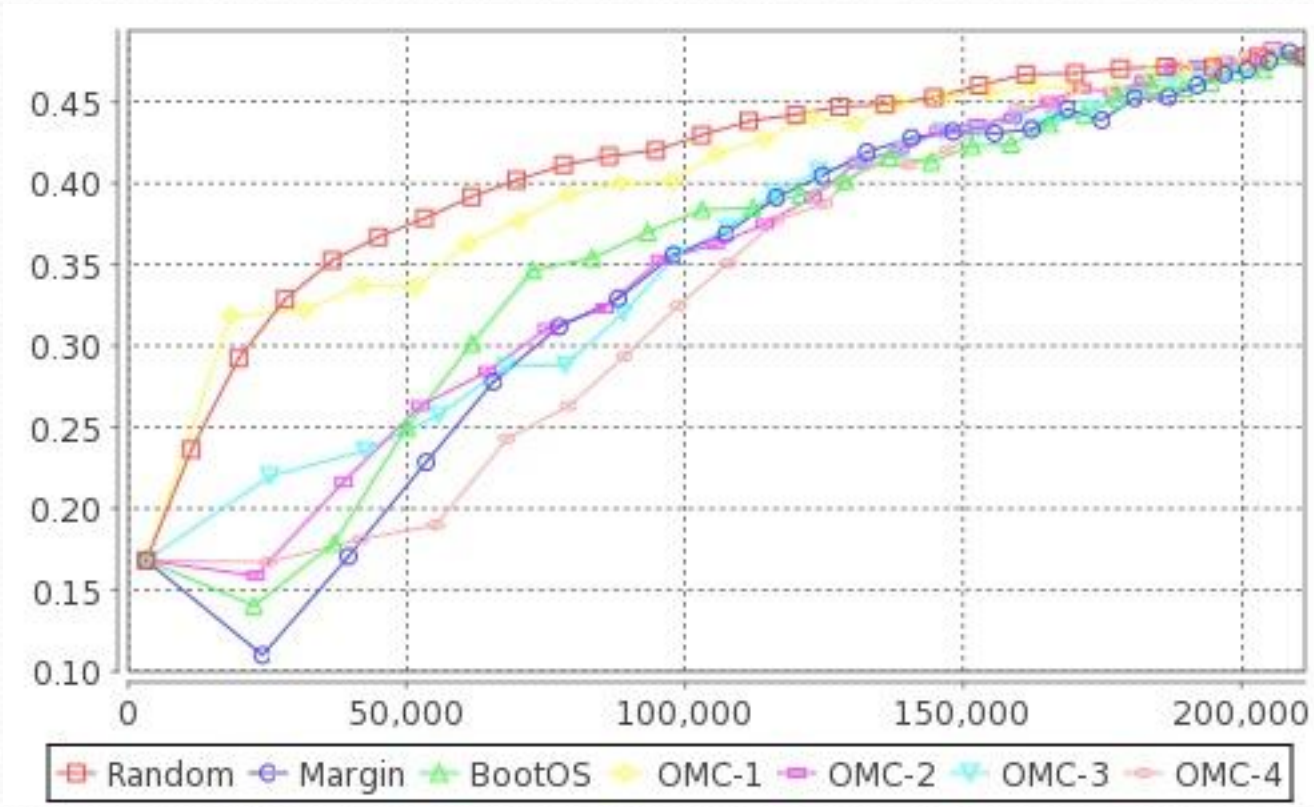
- EAW (Yellow) and EAW-df (Light blue) are among the worst performers
  - the two versions of EAW without using tf-idf

Examining Extension 3 (EAW)



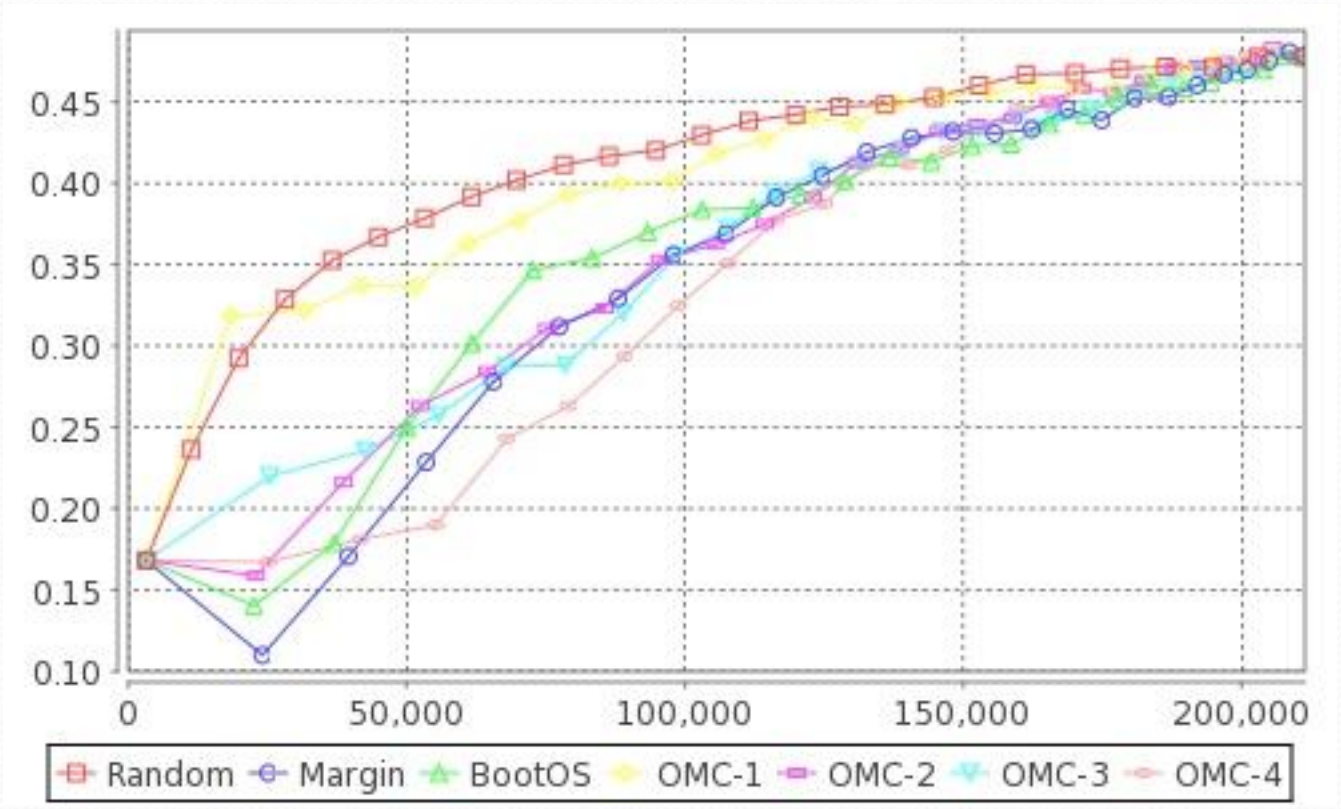
- EAW-tfidf-df (Pink) & EAW-tfidf (Light Pink) are the best performers
  - tfidf is a good measure of term informativeness

Examining Extension 2 (OMC)



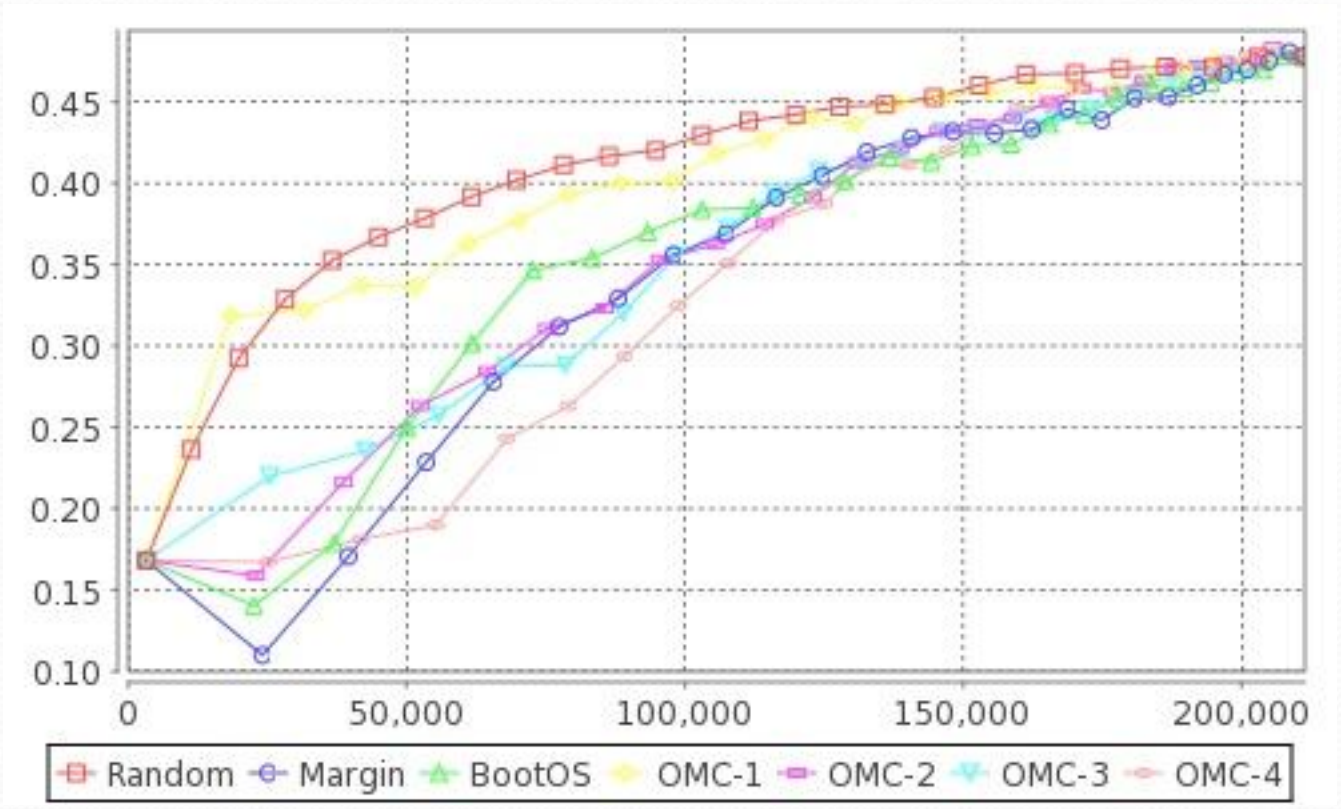
- The OMC extension prefers reports that are informative for multiple classifiers
  - OMC-k: prefers reports that k classifiers are least confident about

Examining Extension 2 (OMC)



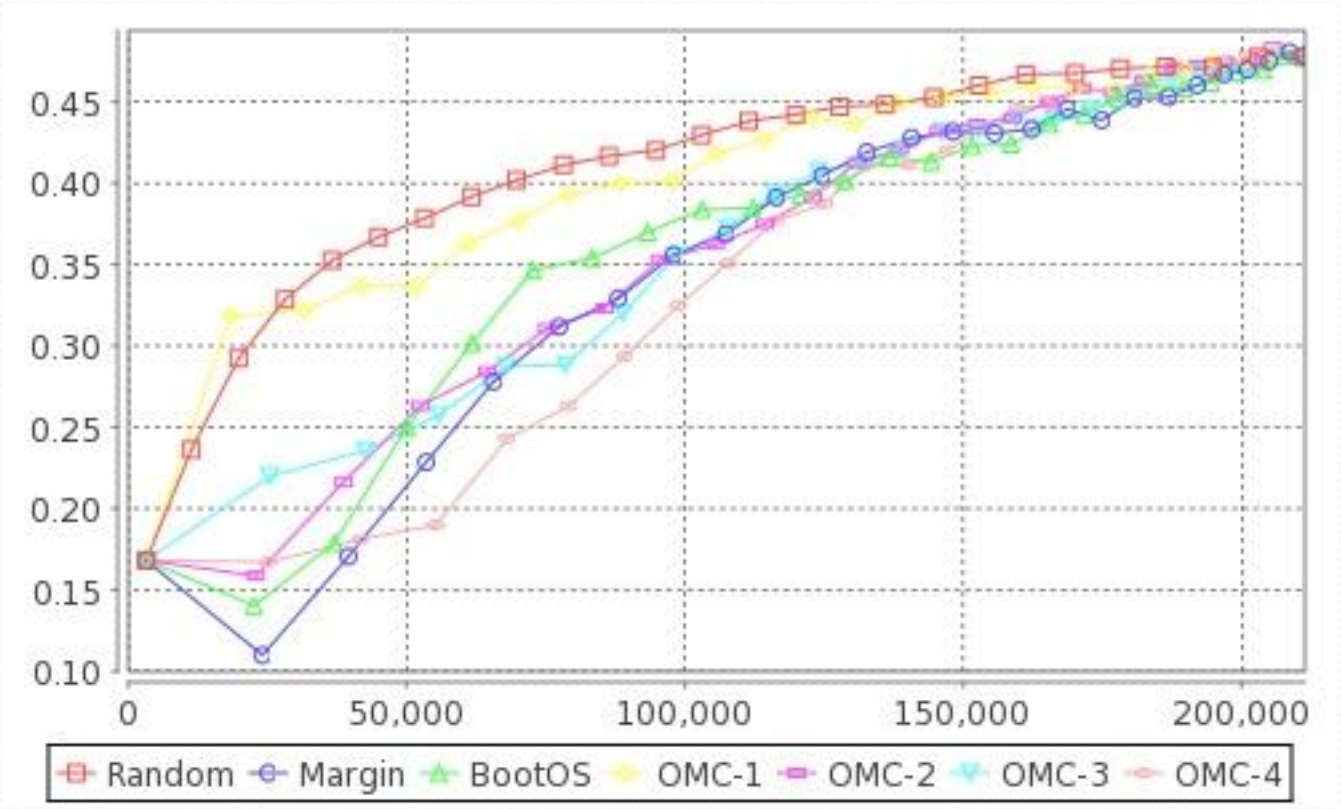
- OMC-1 (Yellow) performs comparably to Random

Examining Extension 2 (OMC)



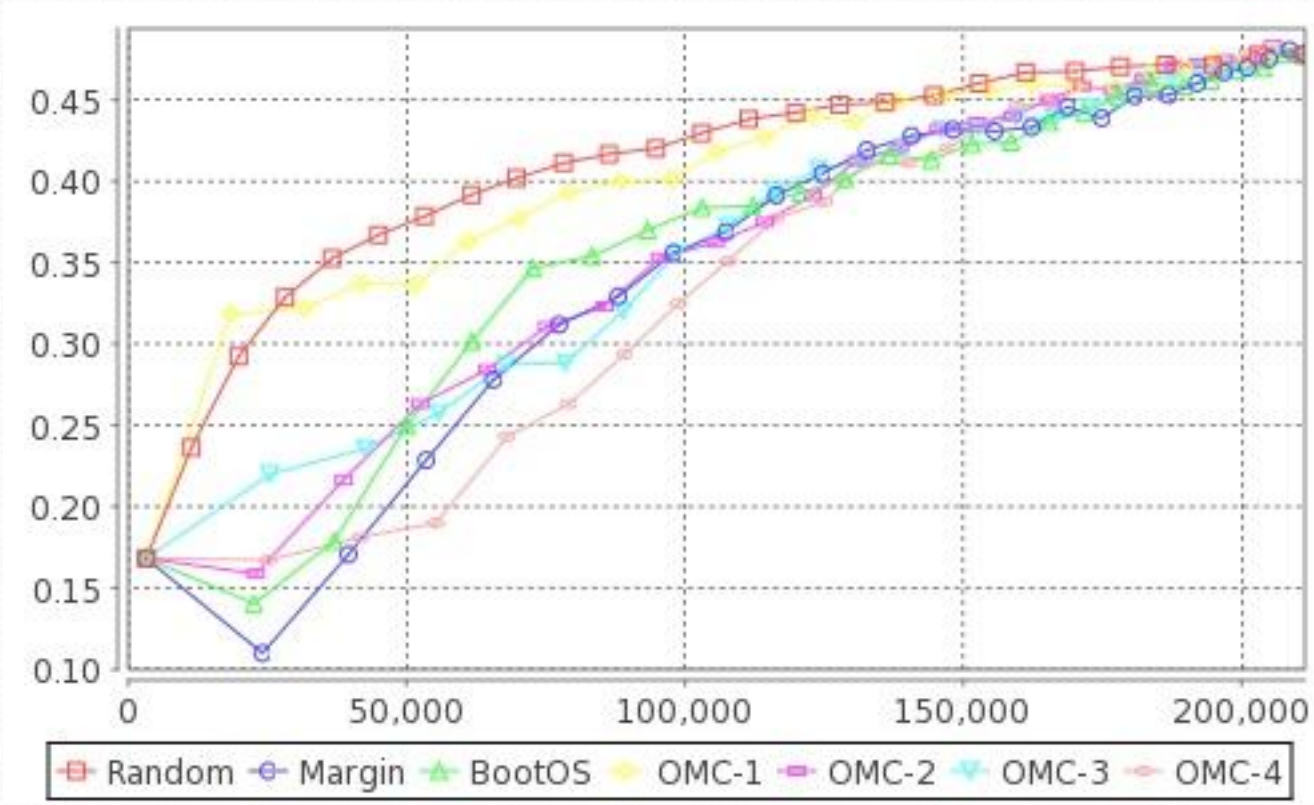
- OMC-2 (Pink), OMC-3 (Light blue), OMC-4 (Light pink) perform poorly
  - Prefer reports that lie close to 2, 3, 4 hyperplanes respectively
  - Problem: select reports that are less close to any hyperplane

Examining Extension 2 (OMC)



- Using only the first two extensions is not effective
  - OMC-1, the best version, performs only comparably to Random

Margin Baseline vs. Random Baseline



- Margin performs worse than Random
  - Margin enforces the “one report per classifier” constraint
    - overly constrains the selection of unlabeled reports

Summary

- Explored and evaluated four extensions to a margin-based active learner for cause identification

- In comparison to the Random baseline
  - the Margin baseline performs worse
  - but the four extensions to Margin yield a reduction in annotation cost for achieving reasonable F-scores by over 50%