Empirical Study of Active Learning Algorithms

Chen Liu, Shi Gao and Ye Tian

University of California, Los Angeles

Motivation

- A whole lot of unlabeled points available, but labels expensive
- Choose data points which are most informative
- Typical scenarios
  - Document (image, video) Labeling
  - Hand-writing recognition (Captcha)
  - Speech recognition
- Goal: accurate classifier with minimum cost

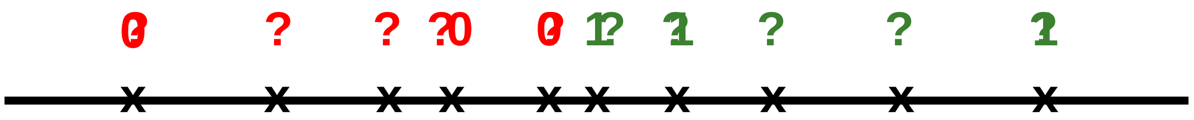
What is Active Learning?

- Active Learning is a subfield of machine learning.
- The key idea of Active Learning is that if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training. [Burr 2010]
- Active learning aims to achieve high accuracy using as few labels as possible, for example

$$poly\left(\frac{1}{\epsilon},\frac{1}{\delta}\right) \rightarrow poly(\log\frac{1}{\epsilon})$$

Active Learning

- Example: threshold
  - Find transition between 0 and 1 labels in minimum steps
- Version space
  - The “possible” hypothesis space according to seen labels
  - Idea: some data points give no additional information to narrow down the version space so we don’t need to learn from it.



Algorithms

- Query by Committee
- Active Perceptron
- Margin Based Active Learning
- These algorithms are efficient and have a proven theoretical error bound under certain assumptions
- We will present
  - The algorithm description
  - Assumptions
  - Theoretical Bounds

Query By Committee

- Description of the algorithm
  - Step 1: Get an unlabeled example $x \in X$ drawn at random from D
  - Step 2: Randomly select two committee members to predict x
  - Step 3: If the two predictions are equal then reject the example and return to Step1
  - Step 4: If the two predictions are different, get label $\mathrm{c}(x)$, set V_n to be all concepts $\mathrm{c}' \in V_{n-1}$ such that $\mathrm{c}'(x) = c(x)$
  - Stop when consecutively reject $\frac{1}{\epsilon} \ln \frac{\pi^2 (n+1)^2}{3\delta}$ examples

Query By Committee

- *Information Gain*
  - The instantaneous information gain from the *i* th label example

$$-\log \frac{Pr_p(V_i)}{Pr_p(V_{i-1})}$$

  - It is proved that there exists a uniform lower bound $1/9 + 7/(18 \ln 2)$ for information gain for any dimension.

Query By Committee

- *Theorem*
If a concept class C has VC-dimension $0 < d < \infty$ and the expected information gain of queries made by QBC is uniformly lower bounded by $g > 0$, then following holds with probability larger than $1 - \delta$,
  - The number of calls to sample is smaller than
  $$m_0 = \max(\frac{4d}{e\delta}, \frac{160(d+1)}{g\epsilon} \max\left(6, \ln \frac{80(d+1)}{\varepsilon\delta^2 g}\right)^2)$$
  - The number of calls to label is smaller than
  $$n_0 = \frac{10(d+1)}{g} \ln \frac{4m_0}{\delta}$$
  - The probability that the prediction algorithm by picking a hypothesis *h* random from version space of QBC makes a mistake is smaller than ϵ. [Freund 97]

Query By Committee

- *Proof*
  - There exists a lower bound for the cumulative information content of first n_0 queries
  - There exists a higher bound for the cumulative information content of the first m_0 examples
  - From first two lemmas, get the relation between m_0 and n_0
  - The number of consecutive rejected examples guarantees that the algorithm stops before testing $m_0 + 1$ examples
  - Gibbs prediction by QBC and consecutive rejected examples gives a error bound of ε

Query By Committee

- For many real-world problems, the committee is infinite.
- The main obstacle in implementing QBC is to sample from the version space (Step 2). It is hard to do this with reasonable computational complexity when d is very large [Ran 2005].
- QBC is very sensitive for noisy data sets.
- We implement original QBC for low dimension data and Active-majority QBC [Liere 97] for high dimension data.
  - Use Winnow algorithm to maintain a finite committee

Active Perceptron

Inputs: Dimensionality d, maximum number of labels L,
and patience R.

$v_1 = x_1 y_1$ for the first example (x_1, y_1).

$\boxed{s_1 = 1/\sqrt{d}}$

For $t = 1$ to L:

Wait for the next example $\boxed{x : |x \cdot v_t| \leq s_t}$ and query its label.

Call this labeled example (x_t, y_t).

If $(x_t \cdot v_t) y_t < 0$, then:

$v_{t+1} = \boxed{v_t - 2(v_t \cdot x_t) x_t}$

$s_{t+1} = s_t$

else:

$v_{t+1} = v_t$

If predictions were correct on R consecutive labeled examples (i.e. $(x_i \cdot v_i) y_i \geq 0 \;\; \forall i \in \{t - R + 1, t - R + 2, \ldots, t\}$), then set $s_{t+1} = s_t / 2$, else $s_{t+1} = s_t$.

[Dasgupta 2005]

Active Perceptron

- Assumptions
  - Data is uniformly distributed on unit ball centered at origin in R^n
  - There exists an oracle

Theorem 3. *With probability* $1-\delta$, *using* $L = O\left(d \log\left(\frac{1}{\epsilon\delta}\right)(\log \frac{d}{\delta} + \log\log \frac{1}{\epsilon})\right)$ *labels and making a total number of errors of* $O\left(d \log\left(\frac{1}{\epsilon\delta}\right)(\log \frac{d}{\delta} + \log\log \frac{1}{\epsilon})\right)$, *the final error of the active modified Perceptron algorithm will be* ϵ, *when run with the above* L *and* $R = O(\log \frac{d}{\delta} + \log\log \frac{1}{\epsilon})$.

Margin Based Active Learning

- Basic idea: choose points with smallest margin to minimize sample complexity.
- If the margin is large than a threshold, the learner reject the point and it will be labeled automatically. Otherwise, the learner query the label and put the point into "working set"
- After enough labels seen, train a new model based on seen labels.
- Repeat the process several iterations and the error rate reduced to ϵ

[Balcan et al. 2007]

Margin Based Active Learning

- ***Realizable Settings***
  - Uniformly distributed on a unit ball in R^d
  - Exists an oracle concept
- ***Parameter Settings***
  - s
  - b
  - C
- ***Bounds***

- ***Algorithm***

Draw m_1 examples into working set
Iterate k = 1…s
  find w_k consistent with all labeled examples in working set
  until m_k points are drawn into ws:
    draw next x
    if $|w_k * x| < b_k$
      put x into ws

Theorem 2. *There exists a constant C s. t. for $d \geq 4$, and for any $\epsilon, \delta > 0$, $\epsilon < 1/4$, using Procedure 2 with $m_k = C\sqrt{\ln(1+k)}\left(d\ln(1+\ln k) + \ln\frac{k}{\delta}\right)$ and $b_k = 2^{1-k}\pi d^{-1/2}\sqrt{5+\ln(1+k)}$, after $s = \lceil \log_2 \frac{1}{\epsilon} \rceil - 2$ iterations, we find a separator of error $\leq \epsilon$ with probability $1-\delta$.*

Margin Based Active Learning

- ***Unrealizable Settings***
  - Uniformly distributed on a unit ball in R^d
  - Satisfies low noise and

$$P_X(|P(Y=1|X) - P(Y=-1|X)| \geq 4\beta) = 1.$$

$$\beta \min\left(1, \frac{4\theta(w,w^*)}{\pi}\right)^{1/(1-\alpha)} \leq \mathrm{err}(w) - \mathrm{err}(w^*)$$

- ***Parameter Settings***

$$\epsilon_k = 2^{-\alpha(k-1)-4}\beta/\sqrt{5+\alpha k \ln 2 - \ln\beta + \ln(1+k)}$$

$$b_k = 2^{-(1-\alpha)k}\pi d^{-1/2}\sqrt{5+\alpha k \ln 2 - \ln\beta + \ln(2+k)}$$

$$m_k = C\epsilon_k^{-2}\left(d + \ln\frac{k}{\delta}\right)$$

$$s = \lceil \log_2(\beta/\epsilon) \rceil$$

- ***Bounds***
  - Excess error

$$\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq 2^{-k}\beta \text{ with probability } 1 - \delta(1 - 1/(k+1))$$

- ***Algorithm***

Draw m_1 examples into working set
Iterate k = 1…s
find w_k \in B(w_$_{k-1}$, r)
clear the working set
until m_k points are drawn into ws:
draw next x
if $|w_k * x| < b_k$
put x into ws

An approximate approach will not looking for w_k in $B(w_{k-1}, r)$, but put unlabeled points into ws with their automatic labels.

Algorithms Recap

Important assumptions:

Algorithms	Linear Separable	Distribution of H	Distribution of X
QBC	Yes	Chosen from a known prior	Chosen from a known distribution on R^d
Active Perceptron	Yes	No	Uniform on unit sphere in R^d
Margin-based Active Learning	No	No	Uniform on unit sphere in R^d

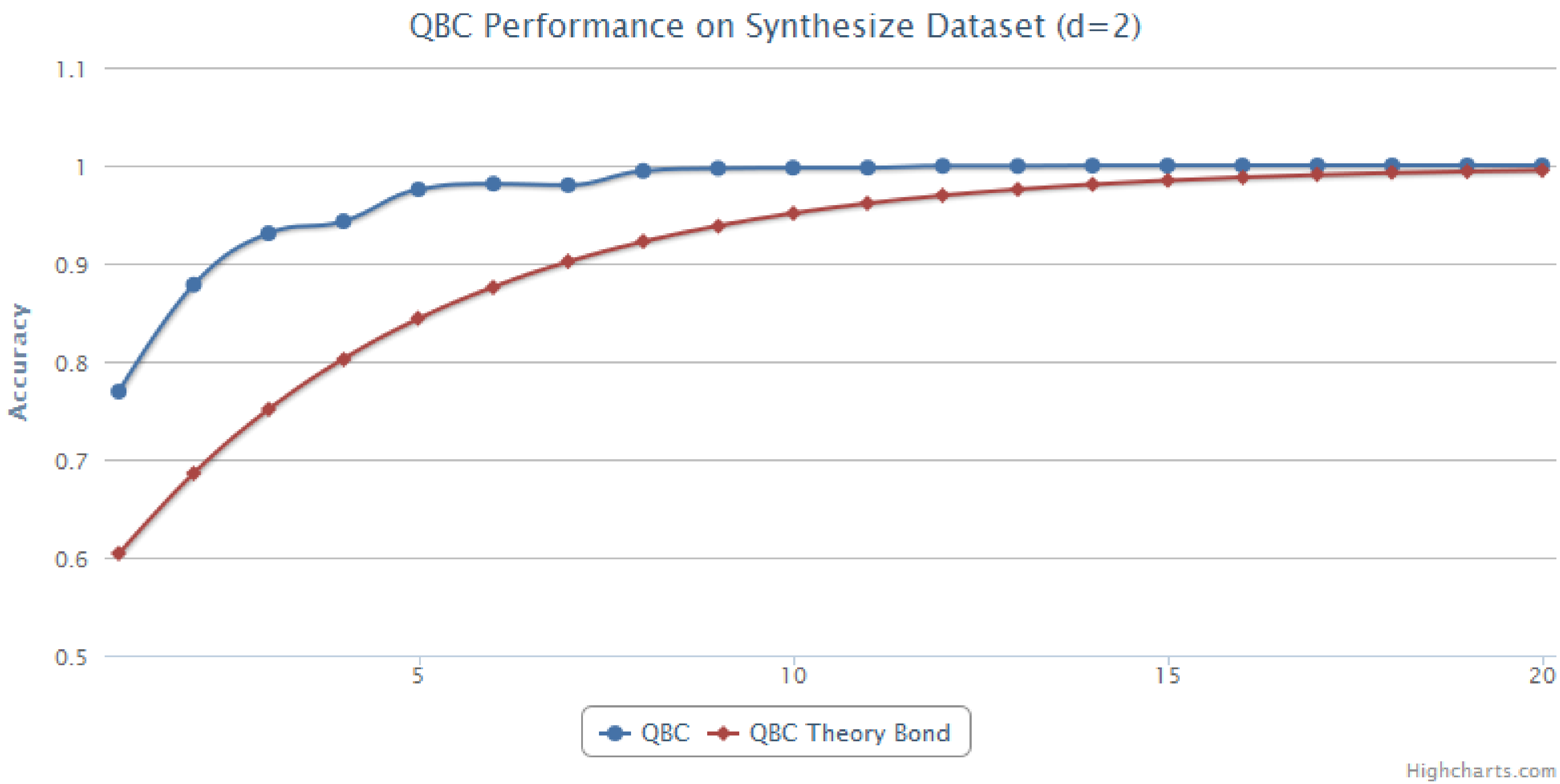
Theoretical bound comparison:

Margin-based < Active Perceptron = QBC

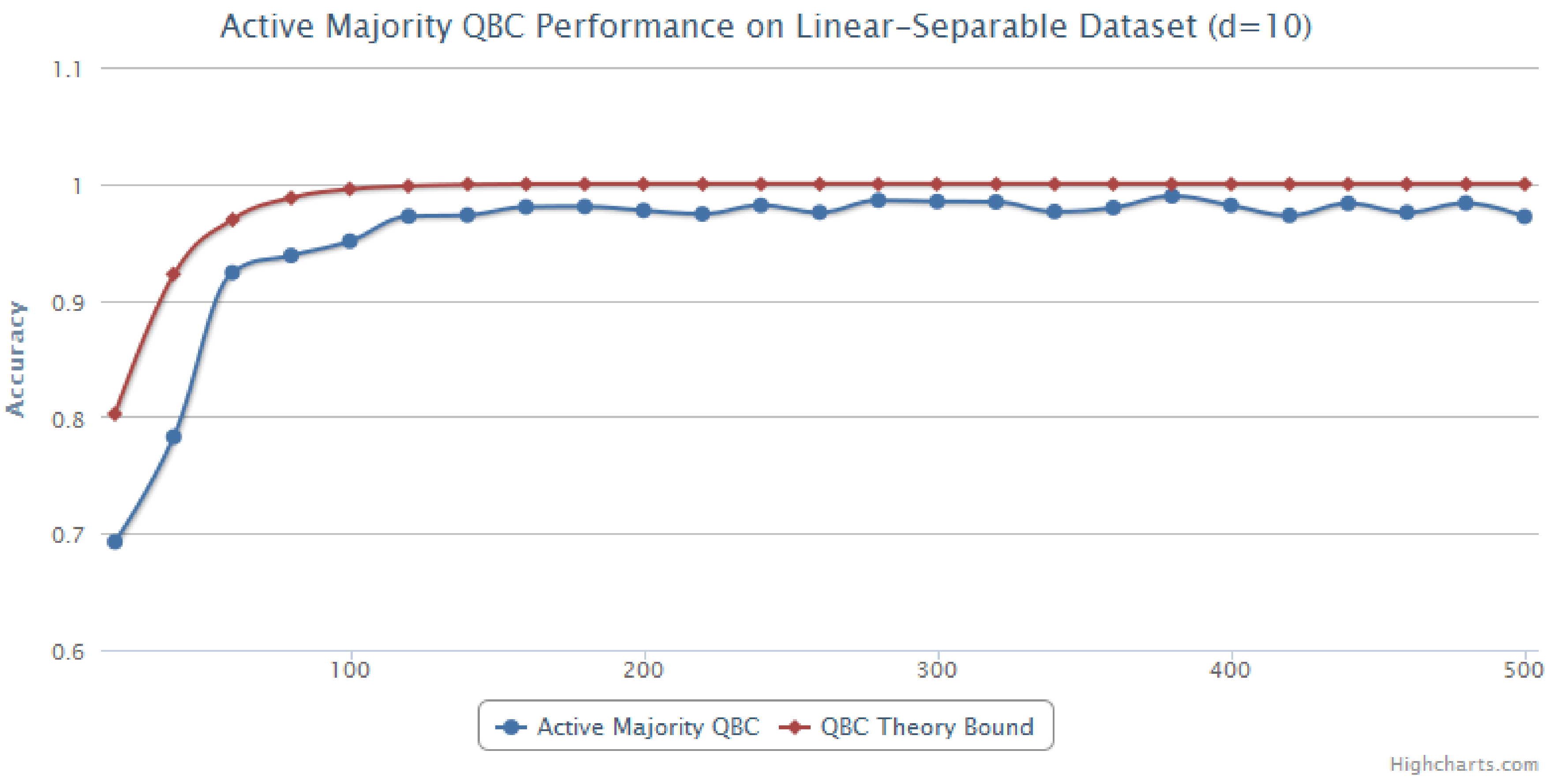
Experiment Setup

- Synthesized Dataset
  - Uniform distributed points on unit sphere centered at origin
  - Exists an oracle linear separator
  - Split into training / testing set:
    - Training points 5000
    - Testing points 3000
  - Compare classification accuracy on testing set
  - Compare with theoretical error bound
  - Compare with baseline method: Perceptron

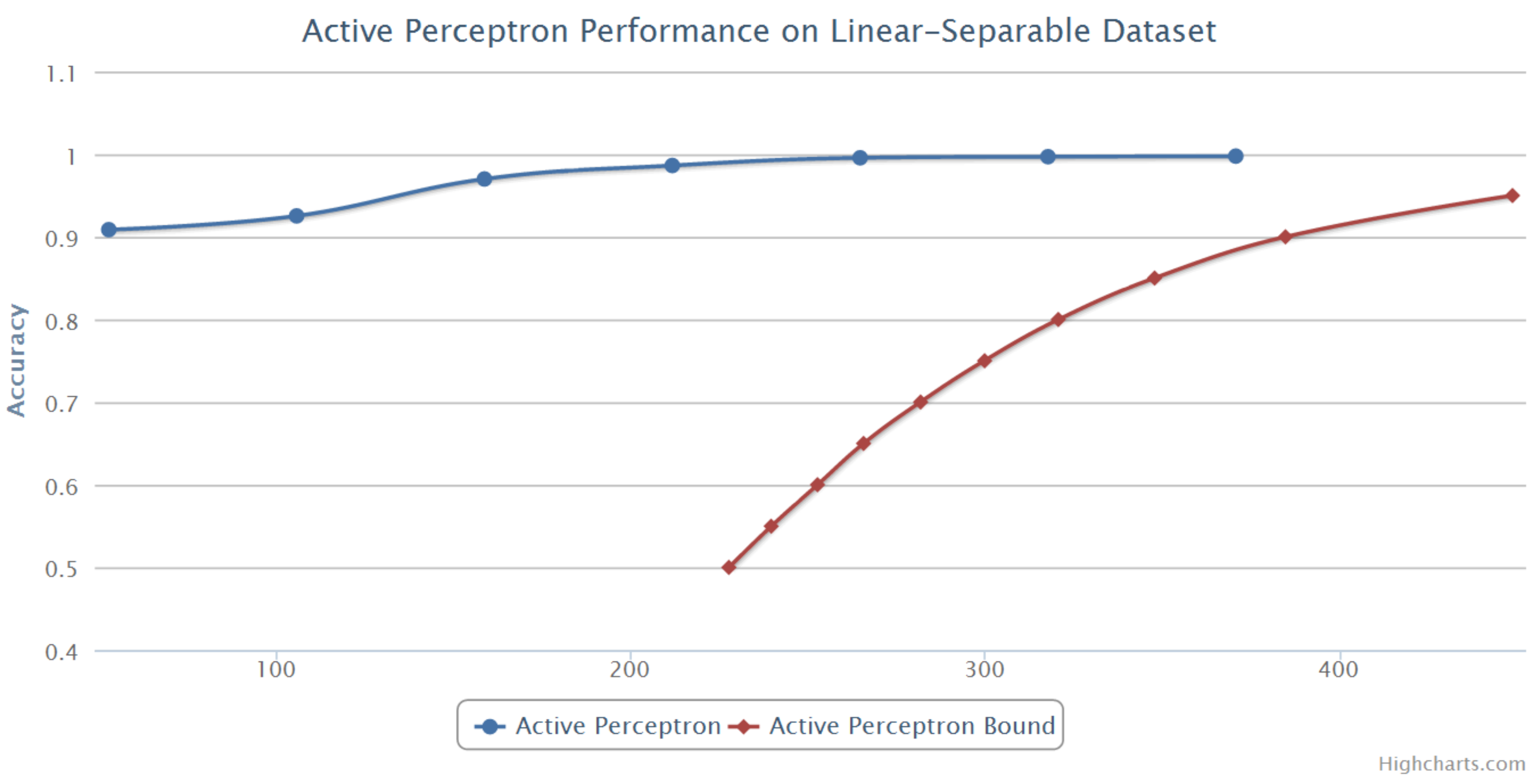
Experiment-Linear Separable



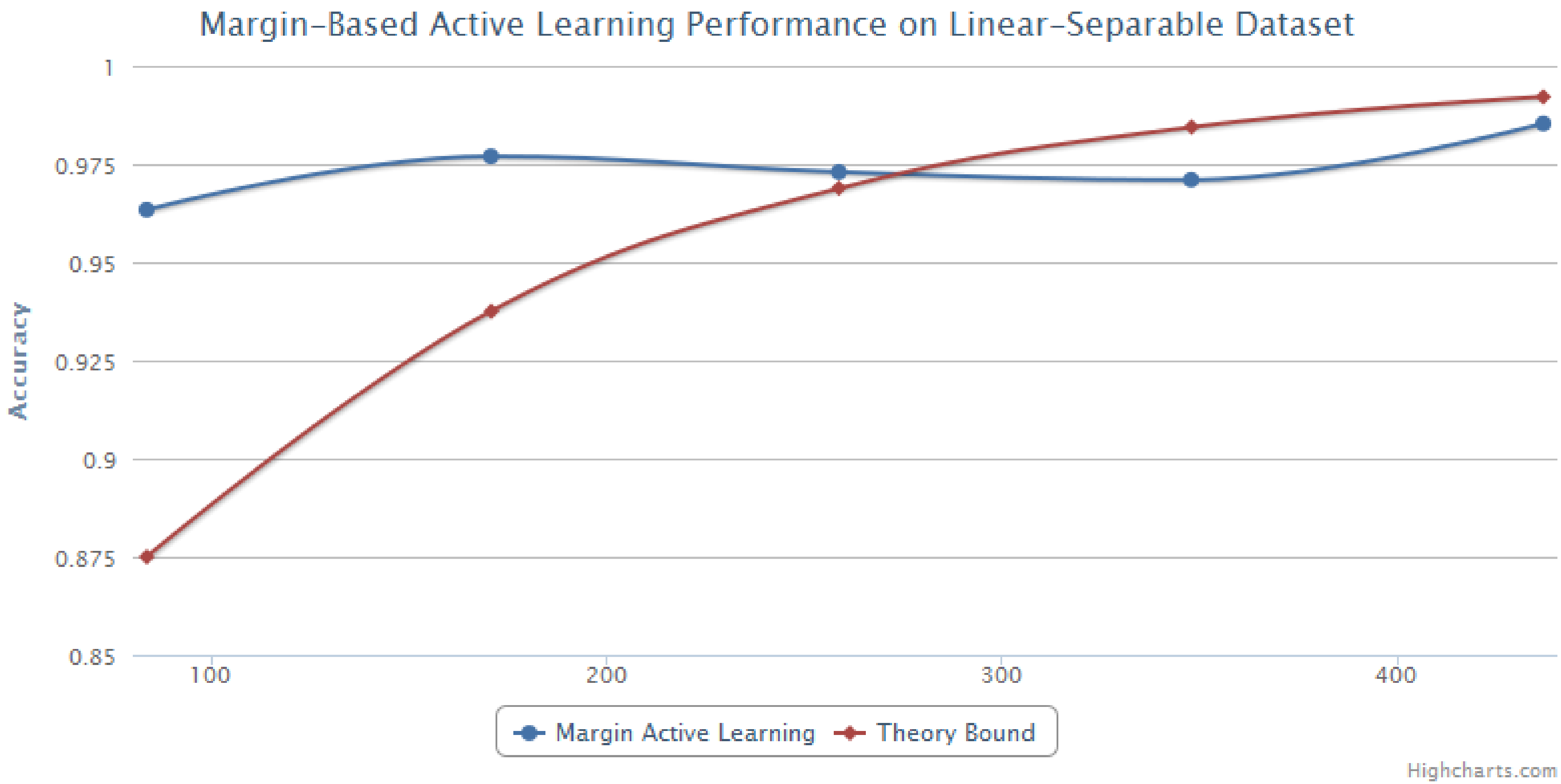
Experiment-Linear Separable



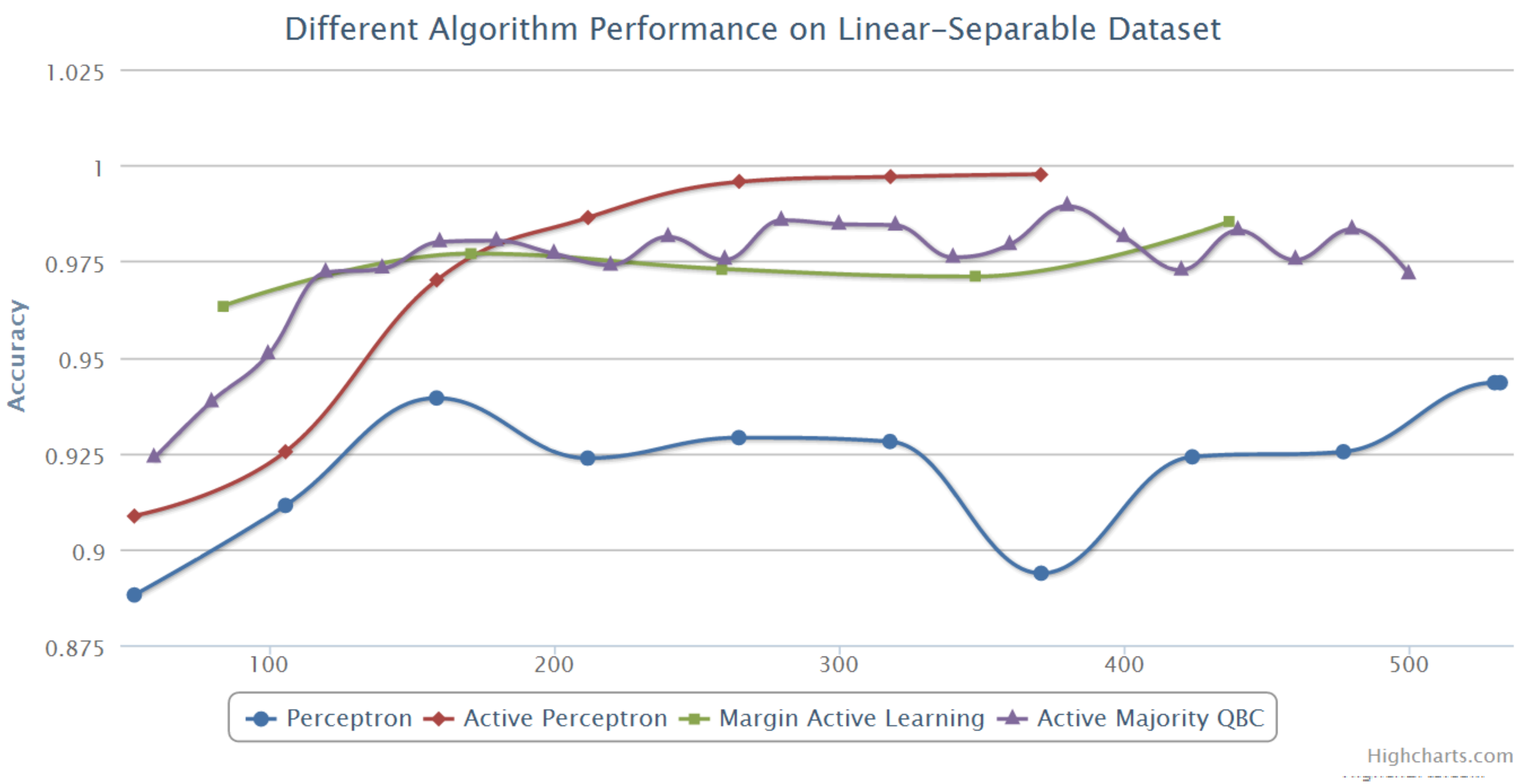
Experiment-Linear Separable



Experiment-Linear Separable



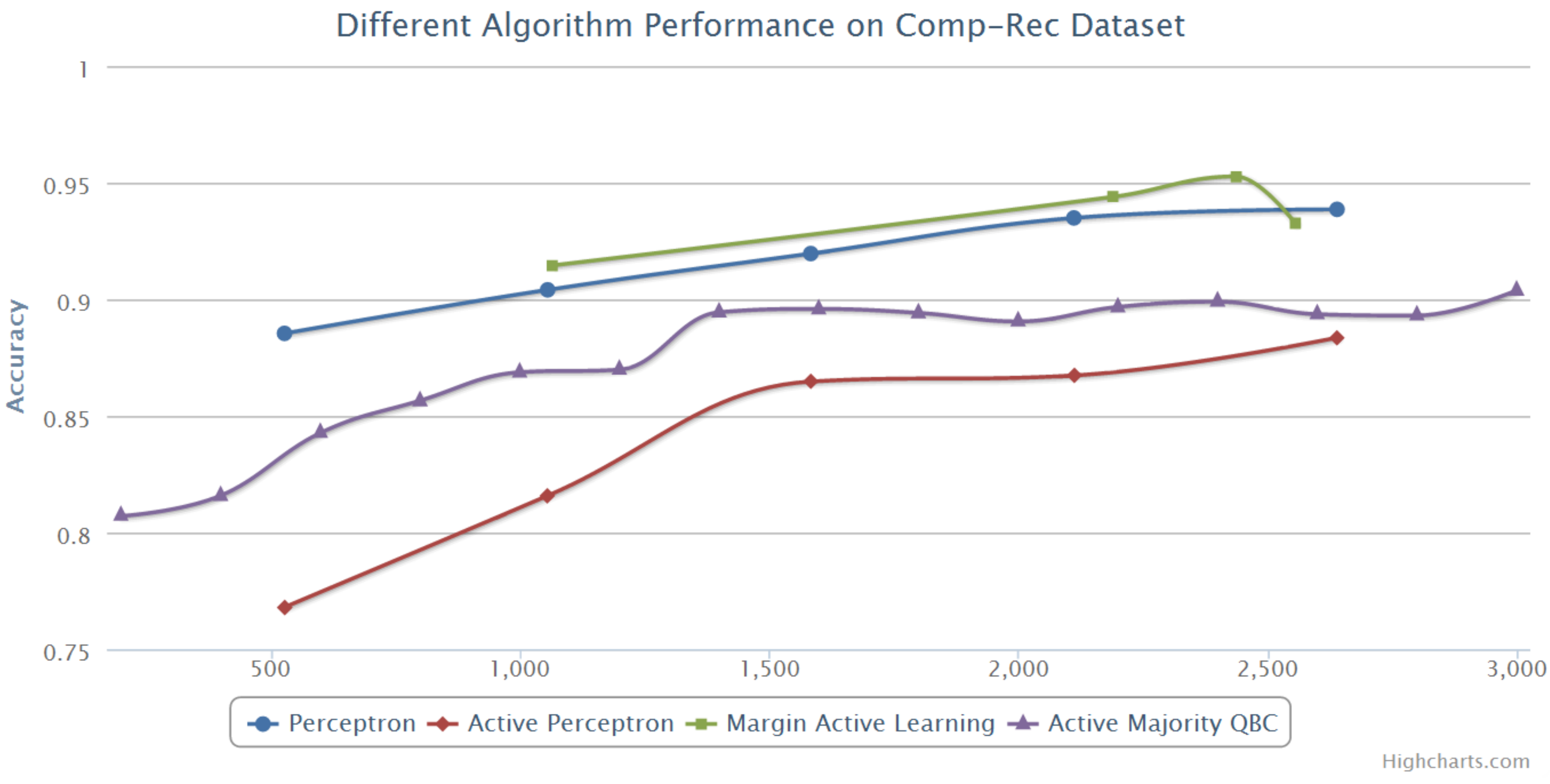
Experiment-Linear Separable



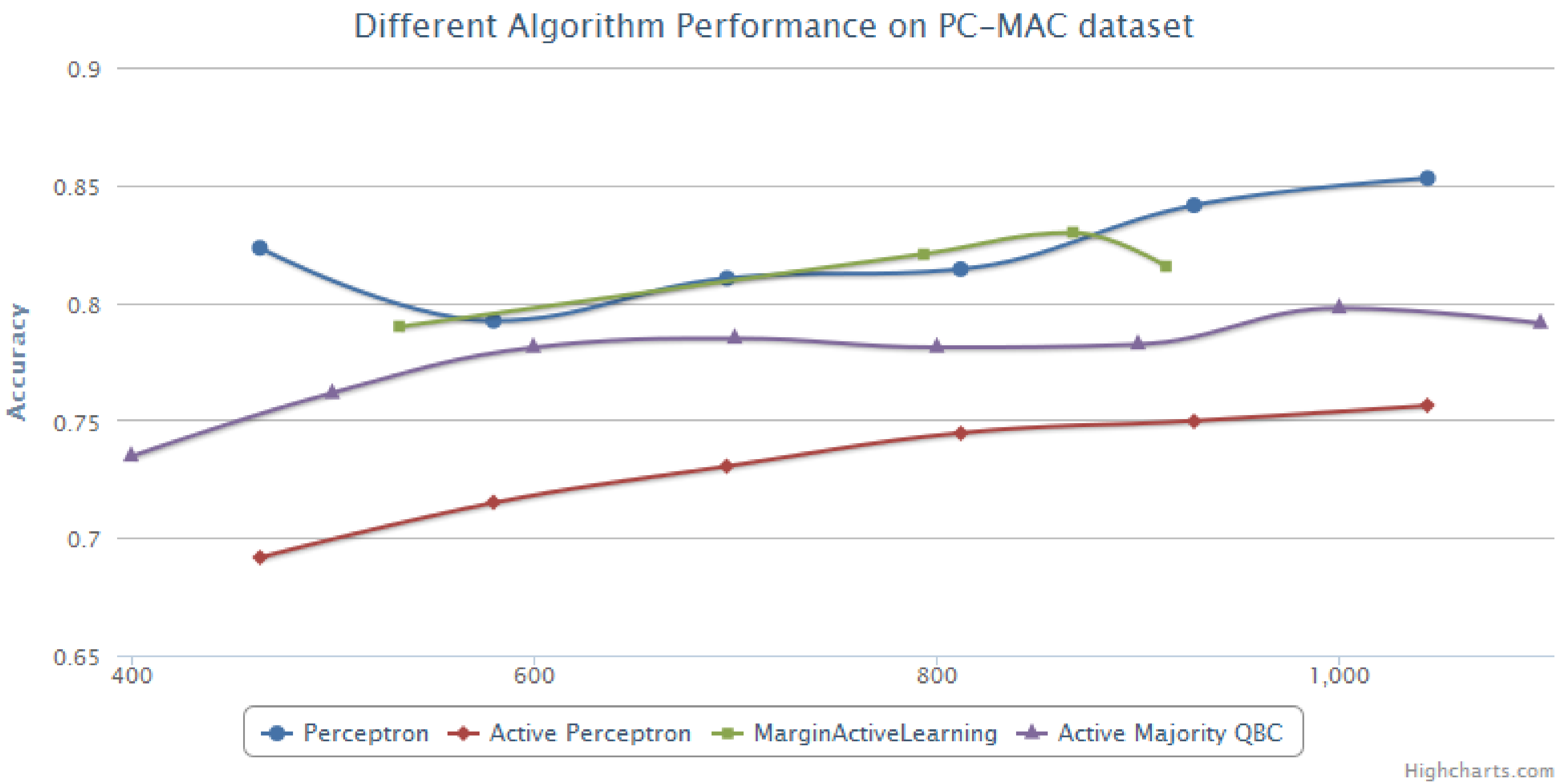
Experiment Setup

- Experiment 2: Real world dataset (20news)
  - Document classification problem
  - Perform three binary classification tasks
    - Recreation vs Computer
    - PC vs Mac
    - Politics vs Religion
  - Learn an linear classifier
  - Feature of the document
    - Normalized tf-idf weighted term vector for each document
- Challenges comparing to synthesis data
  - Assumption on linear separable data won't hold
  - Assumption on concept class distribution won't hold
  - Assumption on data distribution won't hold
  - Very high dimension (60000+ distinct terms)

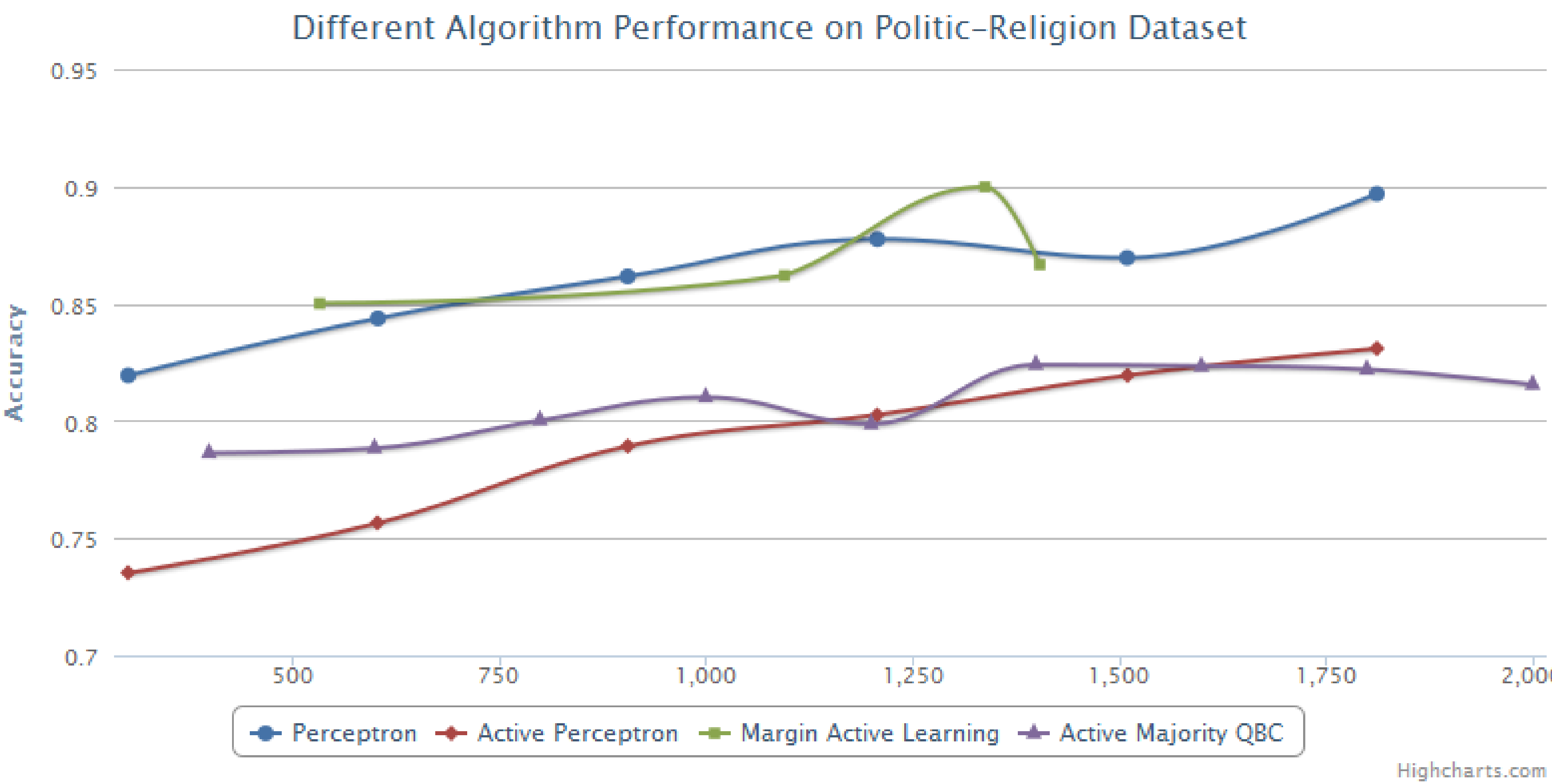
Experiment-Real Data



Experiment-Real Data



Experiment-Real Data



Closing Remarks

- Active learning algorithms performs very well if all assumptions are satisfied.
- However, since the assumptions are hardly satisfied in real world database, the performance gain is not as much as expected.

Reference

[1] Dasgupta, Kalai, and Monteleoni. Analysis of Perceptron-based Active Learning. COLT, 2005.

[2] Freund, Seung, Shamir, and Tishby. Selective Sampling Using the Query by Committee Algorithm. Machine Learning, 1997.

[3] Balcan, Broder and Zhang. Margin-based Active Learning. COLT, 2007.

[4] Active Learning Literature Survey, Burr Settles,Computer Sciences Technical Report, January 26, 2010

[5] Active Learning with Committees for Text Categorization. In proceedings of the Fourteenth National Conference on Artificial Intelligence, 1997

[6] Query by Committee Made Real, Ran Gilad-Bachrach, Amir Navot and Naftali Tishby, NIPS 2005

Q & A