Active Accuracy Estimation on Large Datasets

Sunita Sarawagi, Arun Iyer, Namit Katariya

December 5, 2012

ICDM 2012 Presentation

Outline

Goal

- **Accuracy estimation** : Estimate accuracy of a classifier on a large unlabeled dataset based on a small labeled set and a human labeler
  - **Results** : Between 15% and 62% relative reduction in error compared to existing approaches.
- **Scalable algorithm** : Perform accuracy estimation on unlabeled data so large that it makes even a single sequential scan impractical in an interactive setting
  - **Results** : Close to exact estimates while reading three orders of magnitude less data

Outline

Motivation

- Applications rely on output of classifiers deployed on large data
  - **Examples**: Web page classification, classifying columns to their semantic types
- **Common characteristics**
  - Large and diverse dataset
  - Labeled data unrepresentative of the entire dataset.
- So *Measured accuracy on labeled set* $\neq$ *True accuracy on data*
- Hence need a method that can converge to the true accuracy
  1. An algorithm that returns estimate $\hat{A}$ of true accuracy A of the classifier
  2. *Scalable algorithm* : should work on large datasets where the data is accessible only via an index

Related Work

- Most existing work on *learning* rather than *evaluating* classifiers
- Existing works on selecting instances for evaluating classifiers:
  - (Sawade et al., 2010) present a new proposal distribution for sampling instances
  - (Bennett and Carvalho, 2010) and (Druck and McCallum, 2011) use stratified sampling. However, both assume that classifier $C(\mathbf{x})$ is probabilistic & base their selection on $\Pr(y|\mathbf{x})$ scores
-

Our Solution

- **Stratified sampling**
  - Stratify input space using hash codes (projections on hyperplanes learned over the feature space)
  - Instances in each stratum should have similar accuracy values
  - Estimate accuracy as weighted average of accuracy in each stratum
- **Scaling - Instance selection on large data**
  - TODO: explain where instance selection is needed
  - TODO: describe in a line the idea behind the algorithm to perform accuracy estimation & instance selection on unlabeled data D which can only be accessed via an efficient index partition
- Method agnostic to the type of classifier under consideration

Outline

Results I

Dataset	# Features	Size		Accuracy (%)	
		Seed(L)	Unlabeled(D)	Seed(L)	True(D)
TableAnnote	42	541	11,954,983	56.4	16.5
Spam	1000	5000	350,000	86.4	93.2
DNA	800	100,000	50,000,000	72.2	77.9
HomeGround	66	514	1060	50.4	32.8
HomePredicted	66	8658	13,951,053	83.2	93.9

Table: Summary of Datasets

Results II

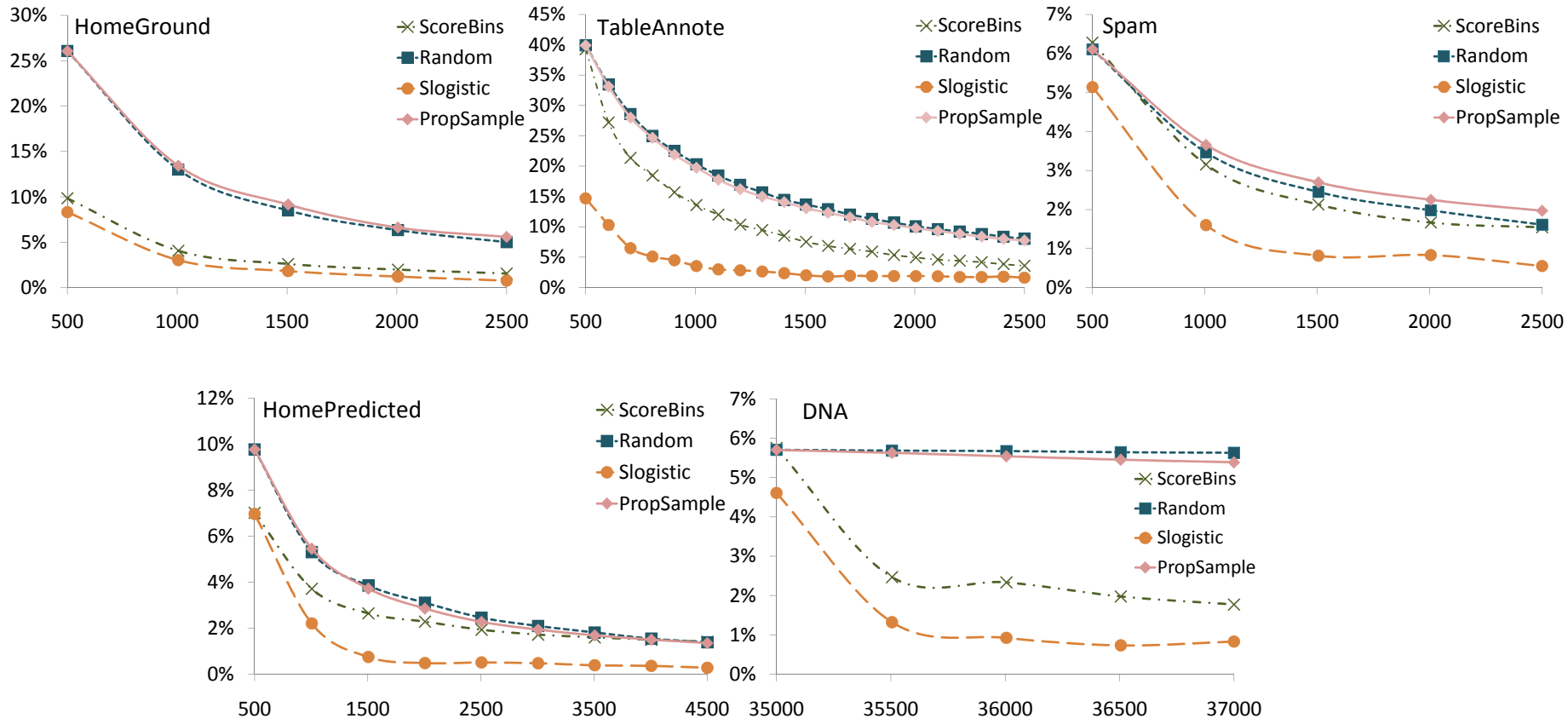

Figure: Absolute error (on the Y axis) of different estimation algorithms against increasing number of labeled instances (on the X axis)

Results III

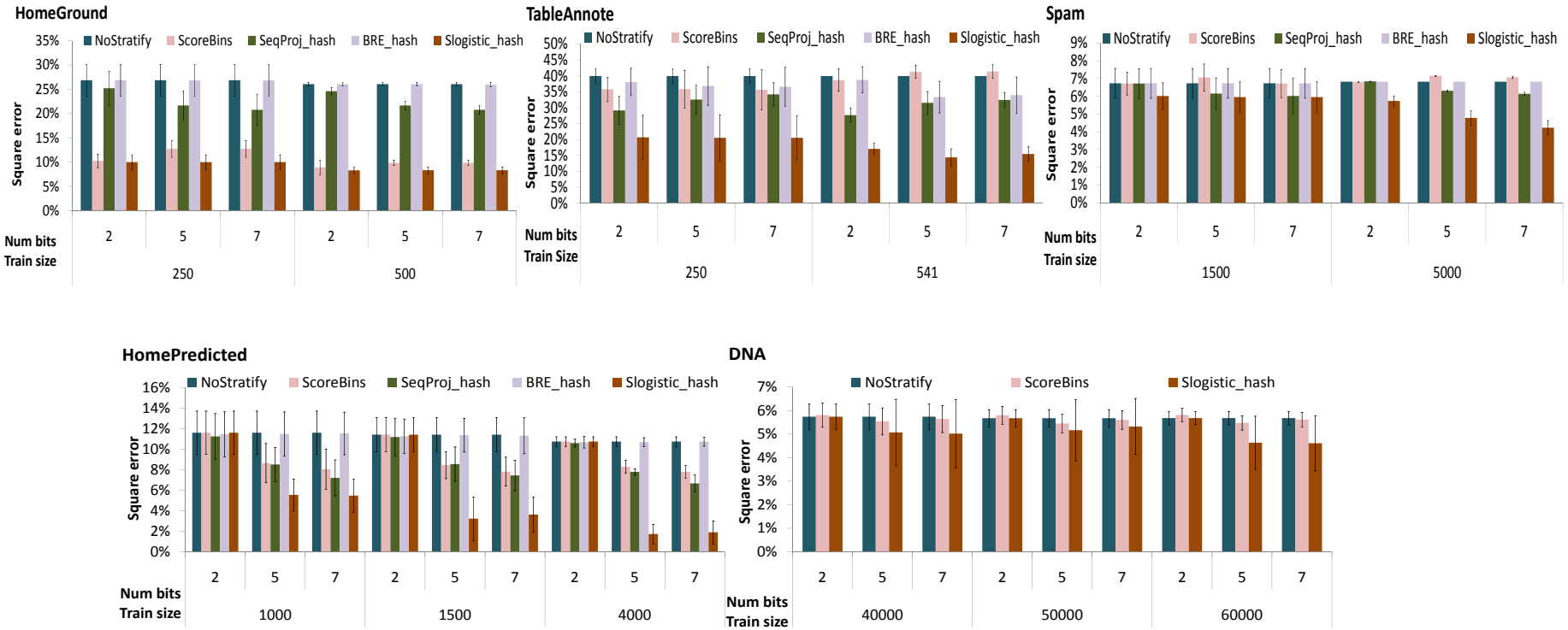

Figure: Error of different stratification methods against increasing training sizes and for different number of bits. The five methods compared: No-stratify, ScoreBins, SeqProj_hash, BRE_hash, Slogistic_hash are presented in this order in each group of bars.

Results IV

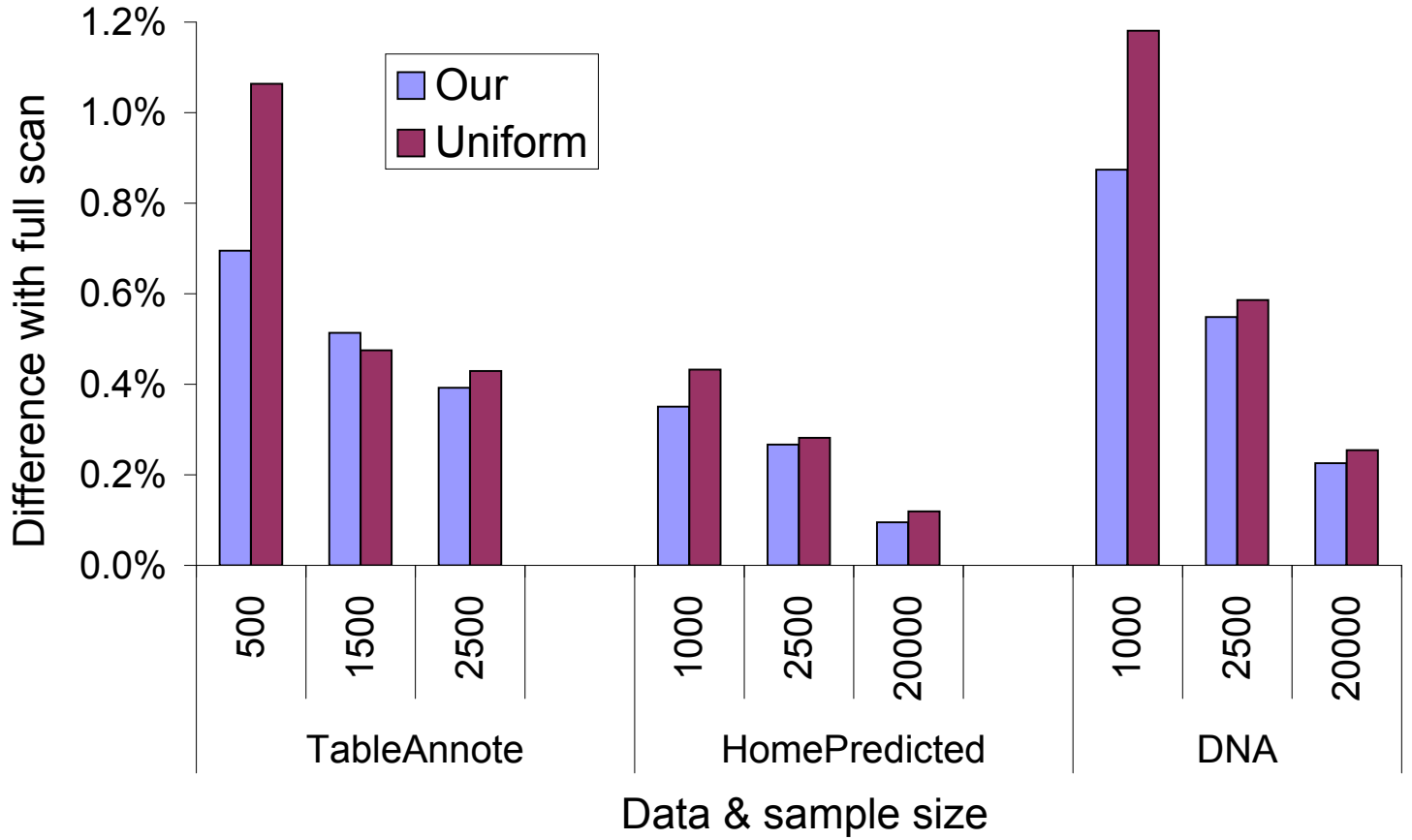

Figure: Comparing methods of sampling from indexed data for estimating bucket weights

Outline

Summary

1. Addressed the challenge of *calibrating a classifiers accuracy on large unlabeled datasets given small labeled data and human labeler*
2. Proposed a stratified sampling based method that provides better estimates than simple averaging & better selection of instances for labeling than random sampling
3. Achieve between 15% and 62% relative reduction in error compared to existing approaches
4. Algorithm made *scalable* by proposing optimal sampling strategies for accessing indexed unlabeled data directly
5. Close to optimal performance while reading three orders of magnitude fewer instances on large datasets

Thank You

References I

Paul N. Bennett and Vitor R. Carvalho. Online stratified sampling: evaluating classifiers at web-scale. In *CIKM*, 2010.

Gregory Druck and Andrew McCallum. Toward interactive training and evaluation. In *CIKM*, 2011.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *ICML*, 2010.