Vox Populi: Collecting High-Quality Labels from a Crowd

Authors: Ofer Dekel, Ohad Shamir (COLT 2009) Presenter: Haofan Zhang

 $\begin{tabular}{ll} \begin{tabular}{ll} David R. Cheriton School of Computer Science, \\ University of Waterloo, \\ \end{tabular}$

Waterloo, Canada

October 20, 2014

Agenda



Introduction

Background Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt Setting the Threshold TReusing the Cleaned Dataset

Experiments

Experiment Settings Experiment Results

Q & A

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

ntroductio

Dackground

Algorithms and Theories

Pruning Can Help

Pruning Can't Hurt

Reusing the Cleaned

Experiments

Experiment Settings Experiment Results



► Traditional machine learning focuses on the single-teacher setting

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Dekel & Shami

Background

Settings and Notations

Theories

Pruning Can Help

Pruning Can't Hurt

Reusing the Cleaned

Experiments

Experiment Settings

Experiment Results

A & C



► Traditional machine learning focuses on the single-teacher setting

▶ We are faced with the problem of learning from crowd

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Lance Control

Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold

Reusing the Cleaned Dataset

Experiments

Experiment Settings

A & C



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Theories

Pruning Can Help

Setting the Threshold T

Dataset

Experiments

Experiment Setting

. . .

- Traditional machine learning focuses on the single-teacher setting
- ▶ We are faced with the problem of learning from crowd
- Therefore, we are interested in identifying and removing low-quality teachers

Introduction

Learning from Crowd



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introducti

Background

Settings and Notations

Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold Reusing the Cleaned

Experiments

Experiments

xperiment Results

2 & A

David R. Cheriton School of Computer Science, University of Waterloo

Several challenges:



Several challenges:

 No prior knowledge on the identity or the quality of the teacher Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

J ---- ---

Theories

Pruning Can Help

Setting the Threshold T

Dataset

Experiments

Experiment Settings

Experiment Results

A & C



Several challenges:

- No prior knowledge on the identity or the quality of the teacher
- No access to gold-set of perfectly labeled examples

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introductio

Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Settings

A & C



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introducti

4 Background

Algorithms and

Theories

Pruning Can't Hurt Setting the Threshold T

Reusing the Cleaned

Experiments

Experiment Setting

ρ. Λ

Several challenges:

- No prior knowledge on the identity or the quality of the teacher
- No access to gold-set of perfectly labeled examples
 - Moreover, a typical teacher only labels a handful set of examples



Vox Populi: Collecting High-Quality Labels from a Crowd

Background

Algorithms and Theories

Experiments

Dekel & Shamir

University of Waterloo

Several challenges:

- No prior knowledge on the identity or the quality of the teacher
- ▶ No access to gold-set of perfectly labeled examples
 - Moreover, a typical teacher only labels a handful set of examples
- ▶ No control on assignment of examples



Vox Populi: Collecting High-Quality Labels from a Crowd

Algorithms and Theories

Experiments

Dekel & Shamir

Background

Several challenges:

- ▶ No prior knowledge on the identity or the quality of the teacher
- ▶ No access to gold-set of perfectly labeled examples
 - Moreover, a typical teacher only labels a handful set of examples
- ▶ No control on assignment of examples
 - Prevent us from applying repeated labeling.
 - ▶ Even applicable, should be avoided because of the cost.

Introduction

Our Goal



Ultimately, our problem is to:

➤ Work with raw labeled data, with single noise.

- Work with raw labeled data, with single noisy label per example
- Detect and eliminate low-quality teaches in a principled and effective manner

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Settings

& A

Introduction

Vox Populi



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

. . .

6 Background

Settings and Notations

Theories

Pruning Can Help Pruning Can't Hurt

Reusing the Cleaned

Experiments

Experiment Settings

ρ. Λ

Vox populi, vox Dei

-The voice of the people [is] the voice of God.



Suppose we have multiple labels for each example

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Dener & Shan

Background

Casiana and Nasasiana

Theories

Pruning Can Help

Setting the Threshold

Reusing the Cleaned Dataset

Experiments

Experiment Setting Experiment Results

2 & A



Suppose we have multiple labels for each example

► If most of teaches are good, we can simply take the average or majority over repeated labels

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T
Reusing the Cleaned
Dataset

Experiments

Experiment Settings Experiment Results

2 & A



Suppose we have multiple labels for each example

- ► If most of teaches are good, we can simply take the average or majority over repeated labels
- ► Then we treat this aggregated label as ground truth and count incorrect label provided by each teacher

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold ${\it T}$ Reusing the Cleaned

Experiments

Experiment Settings

A & C





Suppose we have multiple labels for each example

- ► If most of teaches are good, we can simply take the average or majority over repeated labels
- ► Then we treat this aggregated label as ground truth and count incorrect label provided by each teacher
- ▶ Once we identify low-quality teachers, we can ignore them in the future.

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold

Experiments

Experiment Settings

) P. A



Suppose we have multiple labels for each example

- ▶ If most of teaches are good, we can simply take the average or majority over repeated labels
- ► Then we treat this aggregated label as ground truth and count incorrect label provided by each teacher
- ▶ Once we identify low-quality teachers, we can ignore them in the future.

However, we don't have aggregated labels...

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold 1
Reusing the Cleaned

Experiments

Experiment Settings

O P. A



Suppose we have multiple labels for each example

- If most of teaches are good, we can simply take the average or majority over repeated labels
- ► Then we treat this aggregated label as ground truth and count incorrect label provided by each teacher
- ► Once we identify low-quality teachers, we can ignore them in the future.

However, we don't have aggregated labels...

We want to simulate them!

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold 7
Reusing the Cleaned

Experiments

Experiment Setting

0 0 1



Simulating aggregated labels:

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Theories

Pruning Can't Hurt

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Settings

2 & A



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold 1

Dataset

Experiments

Experiment Settings Experiment Results

A & C

Simulating aggregated labels:

► Specifically, we train a hypothesis(classifier) on the entire unfiltered dataset



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

Sattings and Motations

Algorithms and Theories

Pruning Can Help

Setting the Threshold T

Dataset

Experiments

Experiment Settings Experiment Results

2 A

Simulating aggregated labels:

- Specifically, we train a hypothesis(classifier) on the entire unfiltered dataset
- ► Then we regard the predictions of this hypothesis as the ground truth.



Vox Populi: Collecting High-Quality Labels from a Crowd

Introduction

Background

Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold \it{T} Reusing the Cleaned

Experiments

Dekel & Shamir

Simulating aggregated labels:

- Specifically, we train a hypothesis(classifier) on the entire unfiltered dataset
- ► Then we regard the predictions of this hypothesis as the ground truth.
- ▶ We pretend that we can rely on it, and eliminate low-quality teachers!



We focus on binary classification setting:

▶ Instance space: $\mathcal{X} \subseteq \mathbb{R}^n$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introducti

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Setting Experiment Results





We focus on binary classification setting:

- ▶ Instance space: $\mathcal{X} \subseteq \mathbb{R}^n$
- ► Test Probability Distribution: $\mathcal{D}: \mathcal{X} \times \{-1, +1\}$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T Reusing the Cleaned Dataset

Experiments

Experiment Setting Experiment Results





We focus on binary classification setting:

▶ Instance space: $\mathcal{X} \subseteq \mathbb{R}^n$

► Test Probability Distribution: $\mathcal{D}: \mathcal{X} \times \{-1, +1\}$

Given dataset: $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T
Reusing the Cleaned
Dataset

Experiments

Experiment Setting Experiment Results



We focus on binary classification setting:

- ▶ Instance space: $\mathcal{X} \subseteq \mathbb{R}^n$
- ▶ Test Probability Distribution: $\mathcal{D}: \mathcal{X} \times \{-1, +1\}$
- Given dataset: $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$
- ► The ML algorithm minimizes:

$$\hat{F}_{\lambda}(\mathbf{w}, S) = \lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(f(\mathbf{w}, \mathbf{x}_i), y_i)$$

Additionally,

$$f(\mathbf{w}, \mathbf{x}_i) = \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle$$

represents application of classifier \mathbf{w} to the instance \mathbf{x}_i

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

La La La Contra

Background

Settings and Notations

Theories

Pruning Can't Hurt
Setting the Threshold T

Experiments
Experiment Settings

O & A



In typical supervised learning setting:

lacktriangle we assume that a training set S is sampled i.i.d from ${\cal D}$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold T
Reusing the Cleaned

Experiments

Experiment Settings

Experiment Results

A & C



In typical supervised learning setting:

k teachers:

ightharpoonup we assume that a training set S is sampled i.i.d from $\mathcal D$ Here, we introduce an extra stage where data is labeled by a set of

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

, -----g- ---- ----

Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T
Reusing the Cleaned

Experiments

Experiment Settings

A & C



In typical supervised learning setting:

we assume that a training set \overline{S} is sampled i.i.d from \mathcal{D} . Here, we introduce an extra stage where data is labeled by a set of k teachers:

ightharpoonup There exists k classifiers

$$\{h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_k(\mathbf{x})\} : \mathcal{X} \to \{-1, +1\}$$

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold $\,T\,$ Reusing the Cleaned

Experiments

Experiment Settings Experiment Results



In typical supervised learning setting:

ightharpoonup we assume that a training set S is sampled i.i.d from ${\cal D}$

Here, we introduce an extra stage where data is labeled by a set of k teachers:

- ► There exists k classifiers $\{h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_k(\mathbf{x})\} : \mathcal{X} \to \{-1, +1\}$
- For each unlabeled instance \mathbf{x} , we choose a teacher $t \in \{1, \cdots, k\}$ at random (uniformly here for simplicity)

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Settings and Notations

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold T

Dataset Experiments

Experiment Setting
Experiment Results

O & A



In typical supervised learning setting:

ightharpoonup we assume that a training set S is sampled i.i.d from $\mathcal D$ Here, we introduce an extra stage where data is labeled by a set of k teachers:

- ► There exists k classifiers $\{h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_k(\mathbf{x})\} : \mathcal{X} \to \{-1, +1\}$
- For each unlabeled instance x, we choose a teacher $t \in \{1, \dots, k\}$ at random (uniformly here for simplicity)
- lacktriangle This results in splitting the sample into k subsets, S_1,\cdots,S_k

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Settings and Notations

Theories

Pruning Can't Hurt
Setting the Threshold T

Experiments

Experiment Settings

O & A



This process can be viewed as sampling an unlabeled dataset and labeling it using $\bar{h}(\mathbf{x})$,

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Rackground

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Setting Experiment Results





This process can be viewed as sampling an unlabeled dataset and labeling it using $\bar{h}(\mathbf{x})$, where $\bar{h}(\mathbf{x})$ is the random classifier defined by randomly choosing a hypothesis from h_1, \dots, h_k

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

11 Settings and Notations

Theories

Pruning Can't Hurt
Setting the Threshold T

Experiments

Experiment Setting Experiment Results

A & C





- This process can be viewed as sampling an unlabeled dataset and labeling it using $\bar{h}(\mathbf{x})$, where $\bar{h}(\mathbf{x})$ is the random classifier defined by randomly choosing a hypothesis from h_1, \dots, h_k .
- Remeber we want to minimize $\hat{F}_{\lambda}(\mathbf{w}, S)$:

$$\hat{F}_{\lambda}(\mathbf{w}, S) = \lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(f(\mathbf{w}, \mathbf{x}_i), y_i)$$

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Introduction

Background

Settings and Notations

Algorithms and

Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold T

Experiments

λ & Δ



- This process can be viewed as sampling an unlabeled dataset and labeling it using $\bar{h}(\mathbf{x})$, where $\bar{h}(\mathbf{x})$ is the random classifier defined by randomly choosing a hypothesis from h_1, \dots, h_k .
- Remeber we want to minimize $\hat{F}_{\lambda}(\mathbf{w}, S)$:

$$\hat{F}_{\lambda}(\mathbf{w}, S) = \lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(f(\mathbf{w}, \mathbf{x}_i), y_i)$$

► Then, it can be seen as the empirical counterpart of minimizing

$$F_{\lambda}(\mathbf{w}) = \lambda ||\mathbf{w}||^2 + \mathbb{E}\left[\ell(f(\mathbf{w}, \mathbf{x}), \bar{h}(\mathbf{x}))\right]$$

• We denote \mathbf{w}^* as the minimizer of $F_{\lambda}(\mathbf{w})$

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

Settings and Notations
 Algorithms and

Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold T

Experiments
Experiment Settings

A & G



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Settings and Notations

Experiments

▶ Remember that, our goal is to identify and prune away low-quality teachers.



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Settings and Notations

Experiments

- ▶ Remember that, our goal is to identify and prune away low-quality teachers.
- ▶ After pruning, only a set of high-quality teachers are left



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduction Background

2 Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold TReusing the Cleaned

Experiments

- ► Remember that, our goal is to identify and prune away low-quality teachers.
- After pruning, only a set of high-quality teachers are left
- ▶ We denote $\bar{h}_T(\cdot)$ as randomly pick one classifier from high-quality teachers



► Error rate of teacher *t*:

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold T

Experiments

Experiment Setting

A & C



► Error rate of teacher *t*:

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

▶ Error rate of entire crowd before pruning:

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D} (y\bar{h}(\mathbf{x}) < 0)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold TReusing the Cleaned

Experiments

Experiment Settings

2 A



Error rate of teacher *t*:

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

▶ Error rate of entire crowd before pruning:

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D} (y\bar{h}(\mathbf{x}) < 0)$$

► Error rate of entire crowd after pruning:

$$\bar{e}_T = \Pr_{(\mathbf{x}, y) \sim D}(y\bar{h}_T(\mathbf{x}) < 0|S)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduction

Settings and Notations

Algorithms and Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Settings

Q & A



► Error rate of teacher t:

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

▶ Error rate of entire crowd before pruning:

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D} (y\bar{h}(\mathbf{x}) < 0)$$

► Error rate of entire crowd after pruning:

$$\bar{e}_T = \Pr_{(\mathbf{x}, y) \sim D} (y \bar{h}_T(\mathbf{x}) < 0|S)$$

However, we don't know $\mathcal D$ nor h_t , we cannot calculate e_t directly!

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

_ -----

Background

Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T
Reusing the Cleaned

Experiments
Experiment Settings

0 & A

David R. Cheriton School of Computer Science,



Error rate of teacher *t*:

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

▶ Error rate of entire crowd before pruning:

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D} (y\bar{h}(\mathbf{x}) < 0)$$

► Error rate of entire crowd after pruning:

$$\bar{e}_T = \Pr_{(\mathbf{x}, y) \sim D} (y \bar{h}_T(\mathbf{x}) < 0 | S)$$

However, we don't know $\mathcal D$ nor h_t , we cannot calculate e_t directly! We need to look at something different!

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

Settings and Notations

Theories

Pruning Can Help

Setting the Threshold $\,T\,$ Reusing the Cleaned

Experiments
Experiment Settings

0 & A

David R. Cheriton School of Computer Science,



The true "error-rate" according to \mathcal{D} :

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(y h_t(\mathbf{x}) < 0)$$

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D}(y \bar{h}(\mathbf{x}) < 0)$$

$$\bar{e}_T = \Pr_{(\mathbf{x}, y) \sim D}(y \bar{h}_T(\mathbf{x}) < 0 | S)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notations

Algorithms and Theories

Pruning Can't Hurt

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Settings

Experiment Results

Q & A



The true "error-rate" according to \mathcal{D} :

$$e_t = \Pr_{(\mathbf{x}, y) \sim D}(yh_t(\mathbf{x}) < 0)$$

$$\bar{e} = \Pr_{(\mathbf{x}, y) \sim D}(y\bar{h}(\mathbf{x}) < 0)$$

$$\bar{e}_T = \Pr_{(\mathbf{x}, y) \sim D}(y\bar{h}_T(\mathbf{x}) < 0|S)$$

The idea is to look at the "error-rate" with respect to \mathbf{w}^* :

$$\epsilon_t = \Pr(h_t(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$

$$\bar{\epsilon} = \Pr(\bar{h}(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$

$$\bar{\epsilon}_T = \Pr(\bar{h}_T f(\mathbf{w}^*, \mathbf{x}) < 0 | S)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introductio

14 Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Results

Q & A



The idea is to look at the "error-rate" with respect to \mathbf{w}^* :

$$\epsilon_t = \Pr(h_t(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$
$$\bar{\epsilon} = \Pr(\bar{h}(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$
$$\bar{\epsilon}_T = \Pr(\bar{h}_T f(\mathbf{w}^*, \mathbf{x}) < 0 | S)$$

Also, it is easy to see:

$$\bar{\epsilon} = \frac{\sum_{t=1}^k \epsilon_t}{k}, \quad \bar{\epsilon}_T = \frac{\sum_{t=1}^k \mathbf{1}(t \text{ not pruned})\epsilon_t}{|\{t: \text{t not pruned}\}|}$$

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

5 Settings and Notations

Algorithms and Theories

Pruning Can't Hurt Setting the Threshold

Dataset

Experiments

Experiment Setting

Experiment Results

2 & A



The idea is to look at the "error-rate" with respect to w*:

$$\epsilon_t = \Pr(h_t(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$
$$\bar{\epsilon} = \Pr(\bar{h}(\mathbf{x}) f(\mathbf{w}^*, \mathbf{x}) < 0)$$
$$\bar{\epsilon}_T = \Pr(\bar{h}_T f(\mathbf{w}^*, \mathbf{x}) < 0 | S)$$

Also, it is easy to see:

$$\bar{\epsilon} = \frac{\sum_{t=1}^k \epsilon_t}{k}, \quad \bar{\epsilon}_T = \frac{\sum_{t=1}^k \mathbf{1}(t \text{ not pruned})\epsilon_t}{|\{t: \text{t not pruned}\}|}$$

So, what is the relationship between $\bar{\epsilon}_T$ and classification error?

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

lakas daraktas

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt Setting the Threshold

Experiments

A & C

David R. Cheriton School of Computer Science,

Algorithms and Theoretical Foundations

Relating $\bar{\epsilon}_T$ and Classification Error



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold *T*

Experiments

Experiment Settings

Theorem 1

Assuming $p_D(y|\mathbf{x}) \in \{0,1\}$, it holds for any teacher t that

$$\epsilon_t = \Pr_{(\mathbf{x}, y) \sim D}(yf(\mathbf{w}^*, \mathbf{x}) < 0) + \mathbb{E}_{(\mathbf{x}, y)}[e_t(\mathbf{x}) sign(yf(\mathbf{w}^*, \mathbf{x}))]$$

Algorithms and Theoretical Foundations



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

17 Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold T
Reusing the Cleaned

Experiments

Experiment Settings

. . .

Corollary 2

Assume that for any teacher t, $e_t(\mathbf{x}) \equiv e_t$ is a constant independent of \mathbf{x} . If $\Pr(sign(f(\mathbf{w}^*,\mathbf{x})) \neq y) < 1/2$, then $\{\epsilon_t\}, \bar{\epsilon}, \bar{\epsilon}_T$ are equivalent to $\{e_t\}, \bar{e}, \bar{e}_T$ respectively, up to a uniform, monotonically increasing linear transformation.

Algorithms and Theoretical Foundations



Vox Populi: Collecting High-Quality Labels from a Crowd

Theories

Experiments

Dekel & Shamir

Algorithms and

University of Waterloo

Corollary 2

Assume that for any teacher t, $e_t(\mathbf{x}) \equiv e_t$ is a constant independent of x. If $\Pr(sign(f(\mathbf{w}^*, \mathbf{x})) \neq y) < 1/2$, then $\{\epsilon_t\}, \bar{\epsilon}, \bar{\epsilon}_T$ are equivalent to $\{e_t\}, \bar{e}, \bar{e}_T$ respectively, up to a uniform, monotonically increasing linear transformation.

 \triangleright This means, \mathbf{w}^* does not have to be particularly good, an error-rate smaller than 1/2 suffices.



Motivated by Theorem 1, we consider the following simple algorithm to prune teachers:

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Security and ivolation

neories

Pruning Can Help

Setting the Threshold $\,T\,$ Reusing the Cleaned

Experiments

Experiment Setting

~ ^ ^



Motivated by Theorem 1, we consider the following simple algorithm to prune teachers:

lacktriangle Train a classifier \mathbf{w}' on the entire dataset and prune away any teacher for which

$$\frac{\sum_{i \in S_t} \mathbf{1}(h_t(\mathbf{x}_i) f(\mathbf{w}', \mathbf{x}_i) < 0)}{|S_t|} > T$$

for some threshold $T \in (0,1)$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold TReusing the Cleaned

Experiment Settings
Experiment Popults

A & C



Motivated by Theorem 1, we consider the following simple algorithm to prune teachers:

ightharpoonup Train a classifier \mathbf{w}' on the entire dataset and prune away any teacher for which

$$\frac{\sum_{i \in S_t} \mathbf{1}(h_t(\mathbf{x}_i) f(\mathbf{w}', \mathbf{x}_i) < 0)}{|S_t|} > T$$

for some threshold $T \in (0,1)$

Essentially, this calculates a rough empirical estimate of ϵ_t , and removes all teachers where this estimate exceeds the threshold T

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Dekei & Sham

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold TReusing the Cleaned

Experiment Settings

A & O



Motivated by Theorem 1, we consider the following simple algorithm to prune teachers:

Train a classifier w' on the entire dataset and prune away any teacher for which

$$\frac{\sum_{i \in S_t} \mathbf{1}(h_t(\mathbf{x}_i) f(\mathbf{w}', \mathbf{x}_i) < 0)}{|S_t|} > T$$

for some threshold $T \in (0,1)$

- lacktriangleright Essentially, this calculates a rough empirical estimate of ϵ_t , and removes all teachers where this estimate exceeds the threshold T
- ► The question is: Can this actually help???

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background Settings and Notations

Algorithms and Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold TReusing the Cleaned

Experiment Settings
Experiment Results

A & O

Pruning Can Help



Theorem 3

Assume we use the pruning procedure described previously. Also, let $F:[0,1] \to [0,1]$ be a cumulative distribution function, such that $F(a) = \frac{1}{k} \sum_{t=1}^k \mathbf{1}(\epsilon_t \leq a)$. Let $P \sim F(\cdot)$, and let $N \sim Poi(m/k)$ be a Poisson random variable with parameter m/k. If we assume $m/k = \Theta(1)$ as m,k increase, it holds that

$$\bar{\epsilon} = \mathbb{E}_P[P]$$

and with probability at least $1-\delta$ over the training sample

$$\bar{\epsilon}_T \leq \frac{\mathbb{E}_{P,N}[\Pr(X_N^P \leq NT)P] + r(m,\delta)}{\mathbb{E}_{P,N}[\Pr(X_N^P \leq NT)] - r(m,\delta)}$$

where X_N^P is a binomial random variable, representing sum of N independent Bernoulli random variables with parameter P, and

$$r(m, \delta) = O\left(\sqrt{\frac{\log(6/\delta)}{m}}\right)$$

Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Background

Algorithms and Theories

Pruning Can Help

Setting the Threshold TReusing the Cleaned

Experiment Settings
Experiment Results

A & C

Pruning Can Help



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduction Background

Algorithms and

Pruning Can Help

Setting the Threshold Reusing the Cleaned

Experiments

Pruning Can't Hurt



Another question is:

Can we guarantee that $\bar{\epsilon}_T$ is never considerably larger than $\bar{\epsilon}$?

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduc

Settings and Notation

Theories

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold T Reusing the Cleaned

Experiments

Experiment Setting

A & C

Pruning Can't Hurt



Another question is:

Can we guarantee that $\bar{\epsilon}_T$ is never considerably larger than $\bar{\epsilon}$?

Theorem 4

In the setting of Theorem 3, it holds for any $\{\epsilon_t\}$ that

$$\bar{\epsilon}_T \leq \bar{\epsilon} + \frac{2r(m,\delta)}{\mathbb{E}_{P,N}[\Pr(X_N^P \leq NT)] - r(m,\delta)}$$

where

$$r(m, \delta) = O\left(\sqrt{\frac{\log(6/\delta)}{m}}\right)$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Algorithms and

Theories

Pruning Can't Hurt

Experiments

University of Waterloo

Setting the Threshold T



One final question is, how to choose the threshold T?

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Setting the Threshold T

Experiments

Setting the Threshold T



One final question is, how to choose the threshold T?

Corollary 5

In the setting of Theorem 1, a sufficient condition for $e_{\overline{t}}>\overline{e}$ is

$$\epsilon_t > \Pr_{(\mathbf{x}, y) \sim D} (yf(\mathbf{w}^*, \mathbf{x}) < 0) + \bar{e}$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Natation

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt

Setting the Threshold T

Reusing the Cleaned Dataset

Experiments
Experiment Settings

A & C

Setting the Threshold T



One final question is, how to choose the threshold T?

Corollary 5

In the setting of Theorem 1, a sufficient condition for $e_{t}>ar{e}$ is

$$\epsilon_t > \Pr_{(\mathbf{x}, y) \sim D}(yf(\mathbf{w}^*, \mathbf{x}) < 0) + \bar{e}$$

- This corollary implies that, if ϵ_t is larger than a certain quantity, it is definitely worse than average
- ightharpoonup This suggests a reasonable choice for T is:

$$\Pr_{(\mathbf{x},y)\sim D}(yf(\mathbf{w}',\mathbf{x})<0)+\bar{e}$$

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduction

Background

Settings and Notations

Algorithms and Theories Pruning Can Help

Pruning Can't Hurt

Setting the Threshold T

Reusing the Cleaned Dataset

Experiments

Experiment Settings

Experiment Results

Q & A

David R. Cheriton School of Computer Science,

▶ If pruning is successful, we expect to have a cleaner dataset



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Reusing the Cleaned Dataset

Experiments



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Settings and Notation

Theories

Pruning Can't Hurt

Reusing the Cleaned

Experiments
Experiment Settings

ρ. Λ

▶ If pruning is successful, we expect to have a cleaner dataset

► A more accurate classifier is thereby obtainable



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background

Algorithms and

Pruning Can Help

Setting the Threshold T

Reusing the Cleaned Dataset

Experiment Settings
Experiment Results

. . .

▶ If pruning is successful, we expect to have a cleaner dataset

- A more accurate classifier is thereby obtainable
- ► However, the pruning is data-dependent, therefore generalization will be an issue!

Fortunately, we can address this easily: First, randomly split S into S_1 and S_2



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Reusing the Cleaned Dataset

Experiments



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Fortunately, we can address this easily:

- \triangleright First, randomly split S into S_1 and S_2
- \triangleright Second, we get low-quality teacher set B_1 and B_2 according to S_1 and S_2

Reusing the Cleaned Dataset

Experiments

David R. Cheriton School of University of Waterloo



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Dekel & Slik

Background

Algorithms and Theories Pruning Can Help

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold T

Reusing the Cleaned Dataset

Experiment Settings

^ ^

Fortunately, we can address this easily:

- ightharpoonup First, randomly split S into S_1 and S_2
- ▶ Second, we get low-quality teacher set B_1 and B_2 according to S_1 and S_2
- ▶ Third, use B_1 clean S_2 to get S_2' and use B_2 clean S_1 to get S_1'



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

Algorithms and Theories

Reusing the Cleaned Dataset

Experiments

Fortunately, we can address this easily:

- \triangleright First, randomly split S into S_1 and S_2
- \triangleright Second, we get low-quality teacher set B_1 and B_2 according to S_1 and S_2
- ▶ Third, use B_1 clean S_2 to get S_2' and use B_2 clean S_1 to get S_1'
- Finally, train a classifier on S', where

$$S' = S_1' \cup S_2'$$





Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Theories



25 Experiment Settings





► The data-pruning approach is tested using *Amazon.com's* Mechanical Turk

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

25 Experiment Settings



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

► The data-pruning approach is tested using *Amazon.com's*Mechanical Turk

- ► We create an unlabeled set of over 8,000 examples, each consists
 - ► A search engine query
 - An Internet URL

Introduct

Background

Algorithms and

Pruning Can Help

Pruning Can't Hurt

Setting the Threshold T

Reusing the Cleaned

E.....

Experiments

Experiment Settings

Experiment Results

Q & A





Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background Settings and

Theories

Pruning Can Help

Setting the Threshold TReusing the Cleaned

Reusing the Cleaned Dataset

Experiments

Experiment Settings

Experiment Settings

Experiment Results

O P. A

- The data-pruning approach is tested using Amazon.com's Mechanical Turk
- ▶ We create an unlabeled set of over 8,000 examples, each consists
 - A search engine query
 - An Internet URL
- ► Task was to determine if they are relevant match or not



Vox Populi: Collecting High-Quality Labels from a Crowd Dekel & Shamir

▶ The data-pruning approach is tested using Amazon.com's Mechanical Turk

- ▶ We create an unlabeled set of over 8,000 examples, each consists
 - A search engine query
 - An Internet URL
- ► Task was to determine if they are relevant match or not
- Each example was labeled by 15 different teachers

Algorithms and

Experiments

Experiment Settings

University of Waterloo



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

► The data-pruning approach is tested using *Amazon.com's Mechanical Turk*

- ► We create an unlabeled set of over 8,000 examples, each consists
 - A search engine query
 - An Internet URL
- ► Task was to determine if they are relevant match or not
- ► Each example was labeled by 15 different teachers
- ▶ A total of 375 individual teachers contributed to the dataset

Background

Algorithms and

Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold T

Experiments

Experiment Settings

Q & A



| μ | $\mu=\infty$ (original) | $\mu = 200$ | $\mu = 50$ |
|-------------------------|-------------------------|-------------|------------|
| No. of Teachers | 375 | 881 | 2509 |
| Typical Label / Teacher | NA | 14 | 4 |

Table: Description of 3 Datasets

▶ Parameter μ : each teacher labels at most μ examples

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Experiment Settings





| μ | $\mu=\infty$ (original) | $\mu = 200$ | $\mu = 50$ |
|-------------------------|-------------------------|-------------|------------|
| No. of Teachers | 375 | 881 | 2509 |
| Typical Label / Teacher | NA | 14 | 4 |

Table: Description of 3 Datasets

- Parameter μ : each teacher labels at most μ examples
- ▶ The average of 15 labels are trated as ground truth

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help
Pruning Can't Hurt
Setting the Threshold T

Experiments

Experiment Settings

Q & A



| μ | $\mid \mu = \infty 	ext{(original)} \mid$ | u = 200 | u = 50 |
|-------------------------|---|---------|--------|
| No. of Teachers | 375 | 881 | 2509 |
| Typical Label / Teacher | NA | 14 | 4 |

Table: Description of 3 Datasets

- Parameter μ : each teacher labels at most μ examples
- ▶ The average of 15 labels are trated as ground truth
- ► The training algorithm is well-tuned linear SVM

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Background
Settings and Notations

Algorithms and Theories

Pruning Can Help Pruning Can't Hurt Setting the Threshold 2

Experiments

Experiment Settings

Q & A



Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Theories

27 Experiment Results





Thank you very much!

Any Questions?

Vox Populi: Collecting High-Quality Labels from a Crowd

Dekel & Shamir

Introduc

Settings and Notation

Theories

Pruning Can Help

Setting the Threshold T Reusing the Cleaned Dataset

Experiments
Experiment Settings

28 Q & A