

What if the Irresponsible Teachers Are Dominating? A Method of Training on Samples and Clustering on Teachers*

Shuo Chen, Jianwen Zhang, Guangyun Chen, Changshui Zhang

State Key Laboratory on Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology (TNList)

Department of Automation, Tsinghua University, Beijing 100084, China

{chenshuo07, jw-zhang06, cgy08} @mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn

Abstract

As the Internet-based crowdsourcing services become more and more popular, learning from multiple teachers or sources has received more attention of the researchers in the machine learning area. In this setting, the learning system is dealing with samples and labels provided by multiple teachers, who in common cases, are non-expert. Their labeling styles and behaviors are usually diverse, some of which are even detrimental to the learning system. Thus, simply putting them together and utilizing the algorithms designed for single-teacher scenario would be not only improper, but also damaging. The problem calls for more specific methods. Our work focuses on a case where the teachers are composed of good ones and irresponsible ones. By irresponsible, we mean the teacher who takes the labeling task not seriously and label the sample at random without inspecting the sample itself. This behavior is quite common when the task is not attractive enough and the teacher just wants to finish it as soon as possible. Sometimes, the irresponsible teachers could take a considerable part among all the teachers. If we do not take out their effects, our learning system would be ruined with no doubt. In this paper, we propose a method for picking out the good teachers with promising experimental results. It works even when the irresponsible teachers are dominating in numbers.

Introduction

In most of the previous machine learning literature, the training data is from single teacher or source. In this setting, we deem that the data is i.i.d. sampled from a single hidden distribution. However, this may not be true in many application problems, especially when the Internet has facilitated the acquisition of non-expert opinions from various users nowadays. There have been already a number of crowdsourcing web services for this purposes. For example, the *Amazon Mechanical Turk* (AMT) allows the requesters to publish any “Human Intelligence Tasks” on the website, such as writing essays, filling out certain questionnaires, or just collecting and labeling data. Any user of AMT can finish the tasks he is interested in and get paid.

There are already some examples of using AMT to get labeled data (Sorokin and Forsyth 2008; Deng et al. 2009; Snow et al. 2008). Some other web services include *ESP Game* (Get labels while the users are playing games), *Galaxy Zoo* (Ask the public to classify millions of images of galaxies for NASA) and so on. By aid of them, acquiring non-expert labels is now easy, fast and inexpensive. On the other hand, since there is little control for the teachers, there is no guarantee for labeling quality—there could be careless, fallible, irresponsible or even malicious teachers.

In these systems, the data for learning is collected from multiple teachers, or sources. Each of them does the labeling work according to his own personal style and habit. Therefore, we can think that each teacher represents one joint distribution of samples and labels. Since the label information is usually given by non-expert labelers, the distributions by them could be various comparing to the unknown true distribution. Thus, it would be inappropriate to just simply put them together and manipulate the algorithms designed for single teacher scenario. In recent years, how to make the best use of the labeling information provided by multiple teachers to approximate the hidden true concept has drawn attention of researchers in machine learning.

There has already been some literature for dealing with the multi-teacher setting. One popular strategy is to assign each sample to multiple teachers for labeling, with representing works as (Smyth et al. 1995; Sheng, Provost, and Ipeirotis 2008; Raykar et al. 2009). This repeated labeling strategy has been proved to be useful. However, there are many cases that we have no access in doing so. Even we have, getting multiple labels for one sample could be a great waste of resources. As a result, there is some research on the methods without using repeated labeling. (Crammer, Kearns, and Wortman 2008) makes use of the estimated dissimilarity between teachers as a prior. However, it focuses on finding the most useful samples for training good model for each individual source/teacher rather than training one model for the whole. (Dekel and Shamir 2009b) describes a method for pruning out the low-quality teachers by using the model trained from the entire dataset with all teachers as a ground truth. (Dekel and Shamir 2009a) considers a special case, where the teachers consist of good ones who reveal the true labels and evil ones who deliberately give the reverse labels for binary classification problems. It is based on an

*This work is supported by NSFC (Grant No. 60835002 and No. 60721003).

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

SVM framework, and adds new constraints to reduce the effects by evil teachers. Both of (Dekel and Shamir 2009b) and (Dekel and Shamir 2009a) work well when the majority of data comes from good or high-quality teachers.

We focus on a scenario that is different from previous works in this paper. In real life, when we try to get labels from multiple teachers, we may easily encounter irresponsible ones. By irresponsible teacher, we mean the teacher who disregards the sample itself and gives label as his wish. It is very likely to happen in the services like AMT, where some people just want to make some quick and easy money by labeling in this way. It is also possible when a task is obligatory but not interesting, so some teachers are not willing to take it seriously. Sometimes, the number of irresponsible teachers could be very large, or even dominating in all the teachers. Since their labels cannot be trusted, there should be some method to filter them out. Otherwise, our classifier would be severely damaged. Our work just tries to solve this problem.

The basic idea of our work is that the styles and behaviors of good or high-quality teachers are all very similar, since all of them are trying to convey the unique true concept. They should form a compact cluster in teacher space. On the other hand, each irresponsible teacher's behavior is likely to be different from another irresponsible teacher's, since there is nothing to link them. Also, their behaviors should be different from that of good teachers. We can expect a looser cluster in the teacher space as against the good teachers. Cluster analysis could help to distinguish them. By assuming this, the irresponsible teachers could be recognized even when they take a dominating part.

In this paper, we propose a method for picking good teachers in the scenario mentioned above, while the repeated labeling strategy is not used. We define a sound measurement of the similarity between teachers basing on a k-Nearest-Neighbor classifier a prior, and construct a graph in the teacher space. Then our method finds the good teachers by using spectral clustering for discrimination and semi-supervised learning for refinement. The experimental results are promising. We also show some results of visualizing the teachers' behaviors in this paper, which have not appeared in previous works as far as we know.

Notations and Preliminaries

For simplicity, we focus on binary classification problem in this paper. Suppose we are given a dataset with n samples $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. They are iid sampled from a distribution $p(\mathbf{x}, y)$ defined on $\mathbb{R}^d \times \{+1, -1\}$, with the true label $\{y_1^{true}, y_2^{true}, \dots, y_n^{true}\}$ masked. We assign these samples to m teachers $\{T_1, T_2, \dots, T_m\}$. Each individual teacher T_i gets n_i samples to label, namely $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i}\}$. Since we assume each sample will and will only be labeled by one teacher, we have $\sum_{i=1}^m n_i = n$, and $\{\mathbf{x}_{11}, \mathbf{x}_{12}, \dots, \mathbf{x}_{1n_1}, \dots, \mathbf{x}_{m1}, \mathbf{x}_{m2}, \dots, \mathbf{x}_{mn_m}\}$ is one permutation of $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. The labels given by teacher T_i according to his own style are $\{y_{i1}, y_{i2}, \dots, y_{in_i}\}$, with each $y_{ij} \in \{+1, -1\}$. Since each teacher is characterized by the samples and labels associated with him, we can simply

let $T_i = \{\{\mathbf{x}_{i1}, y_{i1}\}, \{\mathbf{x}_{i2}, y_{i2}\}, \dots, \{\mathbf{x}_{in_i}, y_{in_i}\}\}$ without any misunderstanding.

We assume the assignment of the samples to teachers is random. This means that when $n \rightarrow \infty$, $n_i \rightarrow \infty$, $p_i(\mathbf{x}) \rightarrow p(\mathbf{x})$, $\forall i$, where $p(\mathbf{x})$ is the marginal distribution of $p(\mathbf{x}, y)$ and $p_i(\mathbf{x})$ is the distribution estimated from the infinite set $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i}\}$ by Parzen Windows (Duda, Hart, and Stork 2001). In other words, there should be no difference in the way of allocating samples to different teachers. This assumption ensures that the comparison of different teachers' behaviors is on an equal base.

In this paper, we only consider the case where there are only two kinds of teachers, namely the good teachers and irresponsible teachers. The good teacher is the one who labels each sample $\mathbf{x}$ with its true label y^{true} from the distribution $p(\mathbf{x}, y)$. On the other hand, we denote irresponsible teacher as a teacher T_i who label any sample $+1$ with a fixed probability q_i ($0 \leq q_i \leq 1$) and -1 with $1 - q_i$, no matter what the given sample is.

k-Nearest-Neighbor (kNN) classifier (Duda, Hart, and Stork 2001) is a simple but useful method in the literature of machine learning and pattern recognition. It decides the label of a testing sample according to the voting result of its k nearest neighbors in the training dataset. In fact there is usually no explicit training process, and the training set itself, which is a collection of samples and corresponding labels, can be considered as the classifier. In our settings mentioned above, the allocation of the n training samples to m teachers generates m kNN classifiers naturally, each of which describes the behavior of that teacher. Therefore, we can give the symbol T_i the third meaning in this paper—it stands for the kNN classifier associated with the teacher. For a brief summarization, we use T_i in this paper to denote the i th teacher, the samples and labels he seizes, as well as the kNN classifier ensues when the parameter k is designated.

Proposed Method

In this section, we describe our method of picking a set of good teachers even when the irresponsible teachers are taking a dominating part of all the teachers. We construct a graph to model the teachers in the space, with each teacher as a vertex. We also define the similarity between teachers to link them. Spectral clustering is used to analyze the aggregation of teachers' behaviors, and our assumption of compactness of good teachers serves to pick them out. We also use a semi-supervised learning method to refine the result. We detail each step in the following subsections, and the whole process of the algorithm is listed in Alg. 1.

Similarity between Teachers

Since we want to build a graph on the teachers, we need certain measurement of the similarity between teachers. It is a straightforward idea that if a sufficient large number of samples are labeled by a teacher T_i , we could use them and the teacher's labels to estimate a distribution $p_i(\mathbf{x}, y)$. Then we can use the inner product of functions or Kullback-Leibler divergence to model the similarity or distance between teachers. However, in many cases the amount of

samples is quite limited. When assigned to each teacher, it would be far from enough for estimating a precise distribution to characterize the teacher. As a result, we need to add some a prior to compensate the deficiency of the samples. In this paper, we use a kNN classifier a prior. We deem each teacher as a kNN classifier, which can be used to predict on any samples in $\mathbb{R}^d$. For two teachers T_i and T_j , we use T_i to predict on the samples of T_j , which is $\{\mathbf{x}_{j1}, \mathbf{x}_{j2}, \dots, \mathbf{x}_{jn_j}\}$. We write the predicted label as $\{y_{j1}^i, y_{j2}^i, \dots, y_{jn_j}^i\}$, and the correspondence with the labels given by T_j is

$$c_{i \rightarrow j} = \frac{1}{n_j} \sum_{s=1}^{n_j} \mathbf{1}(y_{js}^i = y_{js}) \quad (1)$$

where $\mathbf{1}(\cdot)$ is the indicator function. It takes the value of 1 if the content in the brackets is true and 0 if otherwise. Then we can define the similarity between teachers T_i and T_j as

$$W_{ij} = \sqrt{c_{i \rightarrow j} c_{j \rightarrow i}} \quad (2)$$

This definition is quite intuitive. The similarity is large when the two teachers agrees with each other's labeling result to great extent, and small if otherwise. It also has two desirable properties. One is that it is symmetric with $W_{ij} = W_{ji}$. This is usually required in any graph-based learning method. The other one is that it is positive within the range from 0 to 1, which is essential in the following clustering process on teachers.

When the number of samples each teacher has goes to infinity, our definition of similarity can be written as

$$\begin{aligned} \lim_{n_i, n_j \rightarrow +\infty} W_{ij} &= \lim_{n_i, n_j \rightarrow +\infty} \sqrt{c_{i \rightarrow j} c_{j \rightarrow i}} \\ &= \sqrt{\int d\mathbf{x} \sum_{y=\pm 1} p_i(y|\mathbf{x}) \sum_{t=0}^{\frac{k-1}{2}} \binom{k-1}{t} \left(p_j(y|\mathbf{x})\right)^{k-t} \left(1 - p_j(y|\mathbf{x})\right)^t} \\ &\quad \cdot \sqrt{\int d\mathbf{x} \sum_{y=\pm 1} p_j(y|\mathbf{x}) \sum_{t=0}^{\frac{k-1}{2}} \binom{k-1}{t} \left(p_i(y|\mathbf{x})\right)^{k-t} \left(1 - p_i(y|\mathbf{x})\right)^t} \quad (3) \end{aligned}$$

When $k = 1$ (Nearest-Neighbor classifier), this expression becomes

$$\lim_{n_i, n_j \rightarrow +\infty} W_{ij} = \int d\mathbf{x} \sum_{y=\pm 1} p_i(y|\mathbf{x}) p_j(y|\mathbf{x}) \quad (4)$$

which is just the inner product in the function space of the teachers' a posteriors. It justifies our definition of similarity from another angle.

As the similarity defined, we build up the structure in the teacher space. We can even visualize by using certain embedding algorithm to map these teachers on a two-dimensional plane. Here in Fig. 1, we use the Kernel PCA technique (Scholkopf, Smola, and Muller 1997) for embedding. It is one example of how good teachers and irresponsible teachers distribute in teacher space. This also enlightens us in the following steps.

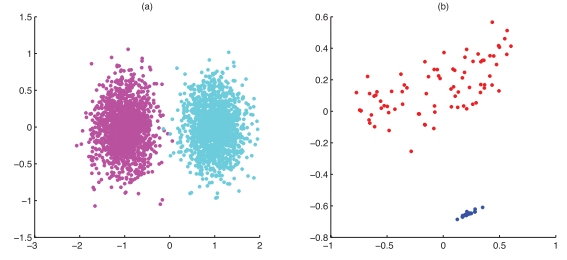


Figure 1: An experiment on toy dataset to show the distribution of teachers. The dataset as depicted in (a) contains 3,000 samples, and is sampled from a Gaussian mixture distribution with two components, each of which stands for one class. We use magenta and cyan to distinguish them. These samples are equally assigned to 100 teachers, of which 20% are good and 80% are irresponsible. By using the similarity we have defined and Kernel PCA, we embed them into a two-dimensional space as shown in (b). The good teachers are represented by blue points and irresponsible teachers by red. We can see that the good teachers are compactly distributed as opposed to the loose cluster of irresponsible teachers.

Spectral Clustering on Teachers

As mentioned in the introduction part, we assume that the good teachers would label the samples with very similar styles, while the styles of irresponsible teachers would be various and usually different from the good teachers'. The example in Fig. 1 just supports our assumption. It is quite straightforward to implement the clustering method to analyze the aggregations of teachers. Since we have the non-negative similarity measurement, we choose to do spectral clustering (von Luxburg 2007) on teachers. We form the similarity matrix W with its element in the i th row and j th column equaling W_{ij} . We also need to reset the diagonal element $W_{ii} = 0, i = 1, 2, \dots, m$. Then the normalized Laplacian matrix is defined as

$$L_n = I - D^{-1/2} W D^{-1/2} \quad (5)$$

where I is the $m \times m$ identity matrix, and D is a diagonal matrix with

$$D_{ii} = \sum_{j=1}^m W_{ij} \quad (6)$$

We can then do the eigenvalue decomposition on L_n to get the two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ corresponding to the two smallest eigenvalues. By using the k-means clustering algorithm (Duda, Hart, and Stork 2001), we can cluster the teachers into two classes if we take each row of $V = [\mathbf{v}_1 \mathbf{v}_2]$ as one data entry. Also, the k-means algorithm could output the average point-to-centroid distance of each cluster, which is a measurement of the compactness of each cluster. We choose the more compact cluster as the coarse set of good teachers we want, and the other cluster for irresponsible teachers. By putting the samples and labels of the two clusters together respectively, we construct two new teachers. They are denoted by T_G and T_I . Since they are

Algorithm 1: Algorithm for picking out good teachers

Input: m teachers $\{T_1, T_2, \dots, T_m\}$. Each $T_i = \{\{\mathbf{x}_{i1}, y_{i1}\}, \{\mathbf{x}_{i2}, y_{i2}\}, \dots, \{\mathbf{x}_{in_i}, y_{in_i}\}\}$;
Output: A set of good teachers;

- 1 **for** $i, j = 1 : m$ **do**
 - Use (1) and (2) to calculate the similarity W_{ij} between T_i and T_j ;
- end**
- 2 Reset the diagonal element of W to 0;
- 3 Use (5) and (6) to calculate the Laplacian Matrix L_n ;
- 4 Calculate the two eigenvectors $V = [\mathbf{v}_1 \ \mathbf{v}_2]$ of the two smallest eigenvalues;
- 5 Use k-means algorithm to cluster rows of V into two classes. The more compact cluster is denoted by T_G , the other by T_I ;
- 6 Use (7) to (10) to calculate the extended similarity matrix W' and Laplacian matrix L'_n ;
- 7 Use (11) to (12) to calculate the good teacher indicator $\mathbf{f}$. Pick out the good teachers according to the sign of $\mathbf{f}$.

the union of several teachers, we can take them as convincing examples of good and irresponsible teacher respectively. They are used in the next step of refinement.

Semi-Supervised Learning for Refinement

This step aims to refine the set of good teachers. We use a semi-supervised learning technique (Chapelle, Schölkopf, and Zien 2006). After the last step, we have two convincing teachers T_G and T_I . They can be considered as labeled teachers, and other teachers $\{T_1, T_2, \dots, T_m\}$ as unlabeled teachers. We first add the two new labeled teachers to our graph, and the similarity matrix of altogether $m+2$ teachers becomes

$$W' = \begin{bmatrix} 0 & W_G^T & W_{GI} \\ W_G & W & W_I \\ W_{GI} & W_I^T & 0 \end{bmatrix} \quad (7)$$

where

$$W_{GI} = \sqrt{c_{G \rightarrow I} c_{I \rightarrow G}} \quad (8)$$

$$W_G = [\sqrt{c_{G \rightarrow 1} c_{1 \rightarrow G}} \quad \sqrt{c_{G \rightarrow 2} c_{2 \rightarrow G}} \quad \dots \quad \sqrt{c_{G \rightarrow m} c_{m \rightarrow G}}]^T \quad (9)$$

$$W_I = [\sqrt{c_{I \rightarrow 1} c_{1 \rightarrow I}} \quad \sqrt{c_{I \rightarrow 2} c_{2 \rightarrow I}} \quad \dots \quad \sqrt{c_{I \rightarrow m} c_{m \rightarrow I}}]^T \quad (10)$$

We can then calculate the normalized Laplacian matrix L'_n over the $m+2$ teachers similarly to (5) and (6). Then the $m \times 1$ class indicator $\mathbf{f}$ is calculated using pseudo-inverse as

$$\mathbf{f} = (L'_n)^\dagger \mathbf{y} \quad (11)$$

where

$$\mathbf{y} = [1 \quad \underbrace{0 \ \dots \ 0}_{m \text{ zeros}} \quad -1]^T \quad (12)$$

The elements of $\mathbf{f}$ that are greater than 0 corresponds the final set of good teachers. Their samples and labels are then used to train a classifier for the task.

Dataset	True Positive Rate	False Positive Rate
<i>COIL</i>	90.72%	9.65%
<i>Digi1</i>	92.73%	7.04%
<i>USPS</i>	91.72%	8.17%
<i>Wine Quality</i>	91.59%	8.25%

Table 1: Average ROC indices for the picked teachers against the preset good teachers on the four datasets.

Experiments

We carry out the experiments on four datasets for binary classification: *COIL*¹, *Digi1*¹, *USPS*¹ and *Wine Quality*.² For preprocessing, we normalize the dataset so that the features of each sample are within 0 to 1. We compare our methods with two other ones. The first one is kNN for single-teachers scenario. We just put all the samples together as the kNN classifier without using any information about the teachers. The second one is Bootstrap Aggregation(Bagging)(Breiman 1996) with kNN, a successful example in the ensemble learning literature. In our setting, this method turns out to be an unweighted vote on all the teachers. For our method, we also use kNN classifier after the good teachers have been picked out.

In the experiments, we let $q_i = 0.5$ for each irresponsible teacher. We vary the ratio of good teachers and the number of samples per teacher n_i for different settings. Then the good and irresponsible teachers are generated accordingly. Each setting is repeated 1000 times to give an average result. Since we are interested in the case when irresponsible teachers are dominating, the ratio of good teachers is kept no more than 0.5 during the experiments. We use 80% of the dataset for training and 20% for testing. The number of nearest neighbors k is tuned by cross validation. The test accuracies on different settings are reported in Fig. 2-5. From the curves we can see that our method is generally better than the two competitors.

For our picking-good-teacher method, we can compare the picked good teachers with the preset good teachers. Thus, we can calculate the average Receiver Operating Characteristic (ROC) indices for each dataset on the teachers. They are listed in Table 1.

Conclusion and Future Work

In this paper, we propose a method for picking out the good teachers from the mixture of good ones and irresponsible ones without using the repeated labeling strategy. Basing on the idea of that the good teachers are all alike as contrary to irresponsible teachers, we make use of clustering analysis and semi-supervised learning to achieve the goal. Also, we define a kNN-classifier-based similarity measurement between different teachers, which can be used for visualizing the teachers' behaviors.

¹<http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html>

²From UCI Machine Learning Repository (Asuncion and Newman 2007). The samples are divided into two class by setting the threshold for rating as 6.

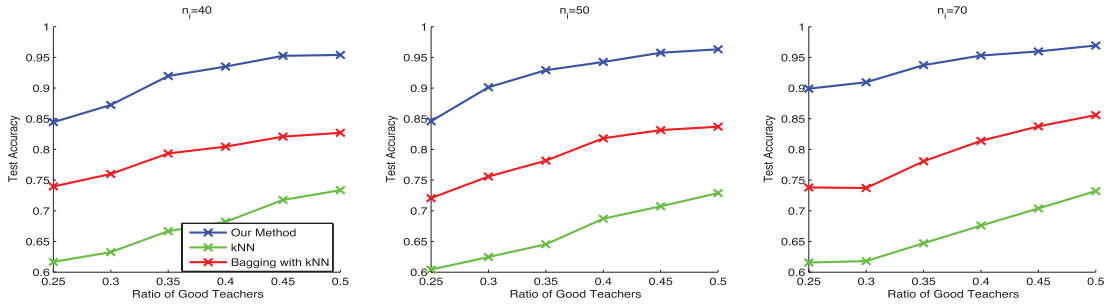


Figure 2: Experimental results on *COIL*.

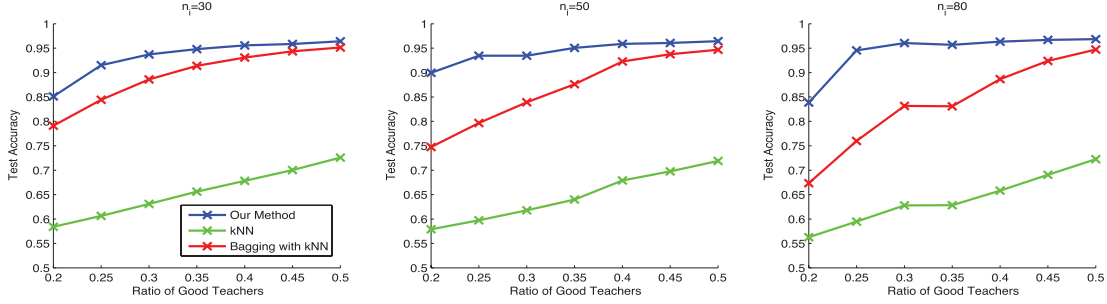


Figure 3: Experimental results on *Digi1*.

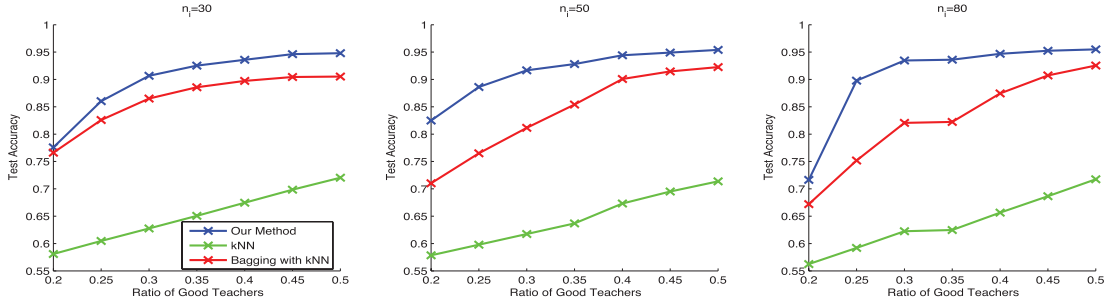


Figure 4: Experimental results on *USPS*.

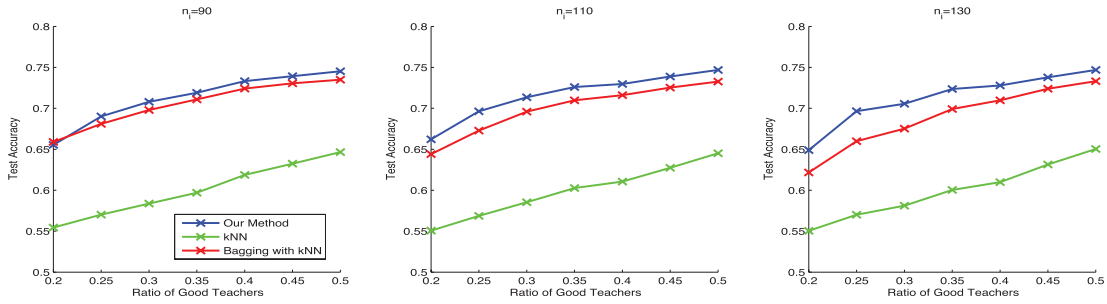


Figure 5: Experimental results on *Wing Quality*.

At the end of this paper, we would like to suggest some possible directions of future research. Since we only focus on a simplified case of mixing irresponsible teachers with good ones in this paper, a lot probable cases have been left untouched. There could be a variety of teachers' behaviors

that end up in complicated distributions in teacher space. We just give two interesting examples during our experiments in Fig. 6. They are rendered by similar methods as in Fig. 1.

In the real life, harshly dichotomizing the teachers into good and irresponsible may not be an accurate enough de-

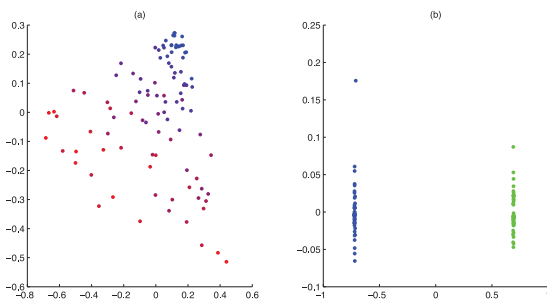


Figure 6: Two examples of different teachers’ behaviors embedded in two-dimensional plane by Kernel PCA. We use the same dataset as in Fig. 1. (a) Teachers with different degrees of responsibility; (b) Good teachers vs. evil teachers.

scription for the behaviors. Also, we may ask: “What kind of behavior is it in between the clusters of the good and the irresponsible?” In Fig. 6(a), we consider the case of teachers with different degrees of responsibility. The degree is the probability that a teacher labels the sample assigned to him seriously. For example, a 0.7-responsible teacher means that he labels his sample as good teacher with probability of 0.7, and as irresponsible teacher with probability of 0.3. In the figure, we use different colors to represent the different degrees of responsibility. The bluer the point is, the more responsible the teacher is. We can see that at one end the teachers with high responsibilities have a tendency of aggregation. At the other end, the teachers with low responsibilities scatter. The transition between the two ends is smooth, just as the transition of colors indicates. This result shows us one possible path that connects the clusters of good teachers and the irresponsible ones. One future direction could be methods that deal with the more diverse behaviors of teachers.

We depict the good teachers vs. evil teachers situation in Fig. 6(b). Here the “evil teacher” is as defined in (Dekel and Shamir 2009a), and we use green point to represent it. We can see that the two kinds of teachers fall into two distinct clusters with similar compactness, thus our methods cannot distinguish them. Moreover, we can expect that a group of teachers would form a compact cluster if they have been arranged to label in the same style, even if this style is far from the true one. This deliberate malicious group action would certainly mislead our learning machine to a wrong concept, and any crowdsourcing should be precautionous about it. A possible solution is adding supervised information. We can let certain trustworthy experts be part of the task, or adding some samples with confident labels to test the teachers. By using the semi-supervised learning technique incorporating these information, we can expect to find the group of good teachers. We believe this is also a very interesting topic for future research.

References

Asuncion, A., and Newman, D. 2007. UCI machine learning repository.

- Breiman, L. 1996. Bagging predictors. *Machine learning* 24(2):123–140.
- Chapelle, O.; Schölkopf, B.; and Zien, A. 2006. *Semi-supervised learning*. MIT press.
- Crammer, K.; Kearns, M.; and Wortman, J. 2008. Learning from multiple sources. *The Journal of Machine Learning Research* 9:1757–1774.
- Dekel, O., and Shamir, O. 2009a. Good learners for evil teachers. In *Proceedings of the 26th ICML*. ACM New York, NY, USA.
- Dekel, O., and Shamir, O. 2009b. Vox populi: Collecting high-quality labels from a crowd. In *Proceedings of the 22nd COLT*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: a large-scale hierarchical image database. In *Proc. CVPR*, 710–719.
- Duda, R.; Hart, P.; and Stork, D. 2001. *Pattern classification*. Wiley New York.
- Raykar, V.; Yu, S.; Zhao, L.; Jerebko, A.; Florin, C.; Valadez, G.; Bogoni, L.; and Moy, L. 2009. Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th ICML*. ACM New York, NY, USA.
- Scholkopf, B.; Smola, A.; and Muller, K. 1997. Kernel principal component analysis. *Lecture notes in computer science* 1327:583–588.
- Sheng, V.; Provost, F.; and Ipeirotis, P. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD*, 614–622. ACM.
- Smyth, P.; Fayyad, U.; Burl, M.; Perona, P.; and Baldi, P. 1995. Inferring ground truth from subjective labelling of Venus images. *Advances in neural information processing systems* 1085–1092.
- Snow, R.; O’Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263. Association for Computational Linguistics.
- Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. *Computer Vision and Pattern Recognition Workshops*.
- von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.