A NEW DATA SELECTION APPROACH FOR SEMI-SUPERVISED ACOUSTIC MODELING

Rong Zhang and Alexander I. Rudnicky

Language Technologies Institute, School of Computer Science
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
{rongz, air@cs.cmu.edu}

ABSTRACT

Current approaches to semi-supervised incremental learning prefer to select unlabeled examples predicted with high confidence for model re-training. However, this strategy can degrade the classification performance rather than improve it. We present an analysis for the reasons of this phenomenon, showing that only relying on high confidence for data selection can lead to an erroneous estimate to the true distribution when the confidence annotator is highly correlated with the classifier in the information they use. We propose a new data selection approach to address this problem and apply it to a variety of applications, including machine learning and speech recognition. Encouraging improvements in recognition accuracy are observed in our experiments.

1. INTRODUCTION

Semi-supervised learning has elicited growing interests in various research fields and many novel approaches have been proposed with promising improvement of performance. Generally speaking, these approaches can be grouped into two categories: generalized EM [1] and incremental learning [2]. Recent research [3] has pointed out that generalized EM may generate a large estimation bias when model assumptions are violated, and consequently deteriorate classification performance. This conclusion is quite discouraging since the situation it describes is very common in speech recognition for which the Gaussian Mixture based acoustic models are only rough approximations to the true underlying distribution.

There is evidence suggesting that semi-supervised incremental learning plus a reasonable data selection strategy, i.e. self-training or co-training, can partially address the degradation problem [4][5]. Same as generalized EM, incremental learning is also performed in an iterative fashion. However, instead of exploiting all of the unlabeled examples, in each iteration of incremental learning only part of them are selected for model training in accordance with some confidence metric. There is a notable characteristic shared by many incremental learning approaches: the selected unlabeled examples must be predicted with high confidence. A first glance gives us the impression that this strategy is reasonable, since a high confidence score usually implies that the corresponding classification result is correct. Expanding the training set with correctly classified examples should therefore improve recognition accuracy. However, counter-examples have been proposed to challenging this concept. For instance, in semi-supervised acoustic model training, [6][7] reported that the best performance is achieved by combining transcribed data with part of un-transcribed data which hypotheses are scored with low confidence.

This paper attempts to investigate this important phenomenon. The analysis we provide in the next section shows that confidence based data selection strategy can also lead to an erroneous estimate of the underlying distribution $P(x,c)$, especially in the case that the confidence annotator is constructed on the information supplied by the classification model. We thus propose a new data selection approach for semi-supervised learning in Section 3, which requires the examples to be selected across the entire feature space complying with $P(x,c)$. To implement the new approach for acoustic model training, we also present a vector conversion scheme that allows K-Means clustering to be performed over the utterance set. Experimental results obtained from a variety of applications will be discussed in Section 4.

2. ANALYSIS ON CONFIDENCE BASED DATA SELECTION

We consider the following scenario of semi-supervised incremental learning in the analysis of data selection approaches. An initial model λ_0 is learned from the labeled set and then applied to classify the unlabeled examples. A confidence metric $f(c;x)$ is used to provide each unlabeled example with a score measuring the likelihood of correctness for the class label given by λ_0. The unlabeled examples with high confidence score are added to the labeled set for training a new model λ_1. The process repeats until all of the unlabeled examples are exhausted or some halting criterion is met.

One example of the confidence metric is the *posterior word probability*, one of the most effective features to estimate the recognition accuracy in continuous speech recognition [8]. Eq. 1 shows its definition in which the hypothesis space is restricted to a word lattice.

$$P_\lambda(< w,t_s,t_e >| x) = \sum_{<w,t_s,t_e>\in h} P_\lambda(h,x) \Big/ \sum_{h\in word\_lattice} P_\lambda(h,x) \quad (1)$$

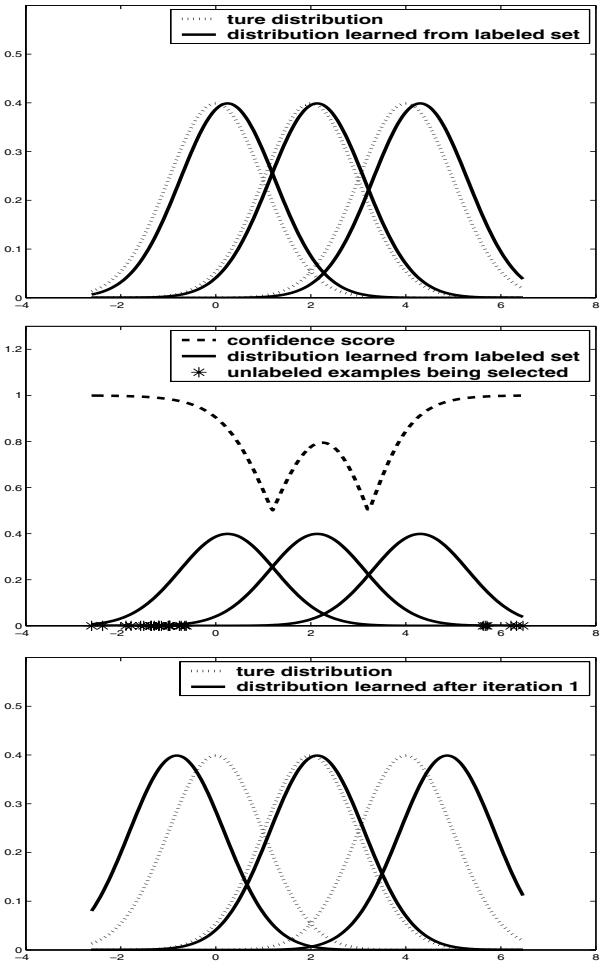
where w is the questioned word with starting time t_s and end time t_e, and h denotes a path in the word lattice.

Posterior word probability demonstrates a common phenomenon in the design of confidence metrics. That is, the confidence annotator is primarily constructed on the basis of the information supplied by the classification model. For example, in Eq. 1, the joint probability $P_\lambda(h,x)$ is calculated from the acoustic score $\log[P_\lambda(x\,|\,h)]$ and language model score $\log[P_\lambda(h)]$ that are provided by the decoding process. Therefore, the selection of unlabeled examples with high confidence score often results in only the examples that match well to the current model being picked, and re-training with such examples will become a process that reinforces what the current model already encodes [6]. Moreover, if the estimation bias exists in the initial classification model, it's likely that the bias will be enhanced rather than eliminated during iteration.

Solely relying on confidence metric for data selection will also lead to other problems. We designed a 1-dimension 3-class classification experiment to illustrate this observation. Suppose the three classes have the same prior probability, that $P(c_1) = P(c_2) = P(c_3)$, and the example x of each class is generated from Gaussian distribution $P(x\,|\,c_i) \sim N(\mu_i,1)$. The mean μ_i

is set to 0, 2 and 4, respectively. To simplify the discussion, we assume the class prior and standard deviation have been known to us. Our task is to learn μ_i for the three classes. In our experiment, a total of 300 examples are generated using $P(c_i)$ and $P(x \mid c_i)$. We randomly pick 10% of them as the labeled set while the remaining as the unlabeled set. Posterior probability $P_\lambda(c \mid x)$ estimated from current model λ is adopted as the confidence metric. In semi-supervised learning, 10% of unlabeled examples with highest confidence score are added to the labeled set for model refining.

Figures 1 to 3 illustrate the learning process. Figure 1 plots the Gaussian density functions trained from labeled examples compared with the true distribution. Due to the scarcity of labeled data, the learned functions obviously deviate from the true concept. In Figure 2, the dashed line depicts the behavior of the posterior probability based confidence metric, while the stars on the bottom denote the unlabeled examples being selected for their high confidence score. Figure 3 shows the new PDF learned from the combination of labeled examples and selected unlabeled examples.



Figures 1~3 Illustration of confidence based data selection

Two important phenomena are revealed in this experiment. First, the unlabeled examples being selected reside only in certain regions of the feature space, i.e. the leftmost and rightmost area as in Figure 2, rather than distribute globally complying with the distribution $P(x)$. Moreover, there is no example picked for some class, i.e. class 2 in this experiment, since most of its examples locate in the area with relative low confidence score. Apparently, learning with the selected unlabeled examples, even though their assigned class labels are correct, will result in an erroneous estimate of the true distribution.

3. NEW DATA SELECTION APPROACH FOR SEMI-SUPERVISED INCREMENTAL LEARNING

The toy experiment described in the previous section suggests that relying solely on confidence metrics for unlabeled data selection is not a risk-free strategy, especially in the case that the confidence annotator is based on the information provided by the classification model. One solution is to employ *external* information not used by classification model to build an independent confidence annotator. Co-training [2] is a successful example of this idea. However, not every real-world application can afford the feature division as required by co-training.

We propose a new data selection principle to address the problems illustrated in the toy experiment: unlabeled examples are not selected across the entire space, and no example is selected for certain class. The approach is shown as follows.

Initialization:

1. Assign class label to each unlabeled example using current model λ.
2. Measure the certainty of each classification with confidence metric $f(c; x)$.

Feature space partition:

3. Partition feature space X into K sub-spaces $D_1, D_2, ..., D_K$ with a reasonable clustering algorithm, i.e. K-means.
4. Let
   n_k^l : number of labeled examples clustered to D_k.
   n_k^u : number of unlabeled examples clustered to D_k.
   $n_{k,c}^l$: number of labeled examples belonging to class c and clustered to D_k.
   $n_{k,c}^u$: number of unlabeled examples classified to class c and clustered to D_k.
5. Compute prior probability for each cluster
   $$P(D_k) = (n_k^l + n_k^u) \Big/ \sum_k (n_k^l + n_k^u) \quad (2)$$
6. Estimate class probability for each class and cluster
   $$P(c \mid D_k) = \frac{\alpha P_l(c \mid D_k) + (1-\alpha) P_u(c \mid D_k) + \beta}{Z} \quad (3)$$
   where
   $$P_l(c \mid D_k) = n_{k,c}^l \big/ n_k^l \quad (4)$$
   $$P_u(c \mid D_k) = n_{k,c}^u \big/ n_k^u \quad (5)$$
   α is a factor balancing $P_l(c \mid D_k)$ and $P_u(c \mid D_k)$, β is a small constant which increases a little bit of the probability of less observed class, and Z is a normalization factor chosen to make $P(c \mid D_k)$ a probability function.

Data Selection:

7. Select cluster D_k according to probability $P(D_k)$.
8. Within cluster D_k, select class c according to probability $P(c \mid D_k)$.

9. For unlabeled examples clustered to D_k and classified as c, add the one with the highest confidence score to the labeled set.
10. Repeat 7 until enough unlabeled examples are selected.

The major difference between the traditional approach and our new approach is that we use not only confidence score as one of the necessary criteria, but also require that the selection comply with the underlying distribution $P(x,c)$, which can be decomposed as $P(x)P(c|x)$. In our approach, $P(x)$, the distribution of examples in feature space, is approximated by $P(D_k)$, the prior probability of clusters, with the assumption that if a cluster is small enough, the examples belonging to it will have the same $P(x)$. On the other hand, the posterior probability $P(c|x)$ is modeled by the probability $P(c|D_k)$, which is estimated by linearly combining $P_l(c|D_k)$ and $P_u(c|D_k)$. Moreover, a smoothing factor β is added to the estimation so that the examples in the class with less observations also have chance to be selected for model re-training.

4. EXPERIMENTS

To test the effectiveness of the proposed data selection approach for semi-supervised incremental learning, we conducted a series of experiments on a variety of applications including classic machine learning problems and continuous speech recognition in a meeting environment.

4.1. Semi-Supervised Multi-Class Classification

Our new approach is first tested on three often referenced benchmark datasets in machine learning research: image-segmentation (*image*), letter-recognition (*letter*), and optical-recognition-of-handwritten-digits (*optdigits*), which can be downloaded from the UCI machine learning repository. The last two datasets have predefined training/test split given in their documents: 16000/4000 examples for *letter* and 3823/1797 examples for *optdigits*. For *image*, we use the first 2000 examples as the training set and the remaining 310 as the test set. The labeled data are further separated from the training set by randomly picking a certain portion of examples along with their labels. Three labeling rates are used in our investigations including 5%, 10% and 20%. To erase the uncertainty caused by random selection, experiments are repeated for 100 times for each labeling rate. The overall means and standard deviation of test accuracy are reported as the final performance. Our experiments use Gaussian Mixtures as the base classifier, K-Means as the clustering method, and Negative Entropy [9][10] as the confidence metric which definition is as follows.

$$ne(x) = \sum_{m=1}^{M} P_\lambda(c_m|x)\log P_\lambda(c_m|x) \quad (6)$$

Our new data selection approach is then compared with three other learning methods: (1) supervised training on the labeled set, (2) semi-supervised EM using all the unlabeled examples, and (3) traditional incremental learning that always selects high confidence data. Experimental results are reported in Table 1, in which the best result for each dataset and labeling rate is marked in boldface type. Our new approach works very well in the experiments, which consistently performs better than or as well as the best of the other three methods.

4.2. Semi-Supervised Meeting Recognition

We also applied the new data selection approach to semi-supervised acoustic modeling for the ICSI meeting [CMU1]domain [11]. The dataset has a total of 75 meetings, accounting for 60 hours of raw speech data. We use 10 meetings as the labeled set for initial acoustic training, 61 meetings as the unlabeled set for semi-supervised learning, 3 meetings as the hold-out set for recognizer tuning, and 1 meeting (containing about 7500 words[CMU2]) as the test set. A 13-dimension MFCC feature vector is computed for each frame and then expanded to 39-dimension by adding delta and delta-delta coefficients. The phone set contains 49 basic phonemes. In context dependent training stage, these phonemes are transformed to triphones and then tied together to make 2000 senones. Each senone is modeled using a mixture of 32 Gaussians, giving a total of 64K Gaussians for acoustic modeling.

We employ a neural network based confidence annotator to measure the correctness of hypothesis. The inputs to the neural network consist of four features representing both language model and acoustic model information: *LM-backoff-mode*, *posterior-utterance-probability*, *posterior-word-probability* and *posterior-frame-probability*, while the output is trained to approximate the word accuracy of each hypothesis. We previously found [7] [CMU3]that the confidence score given to the questioned hypothesis is generally proportional to its word accuracy; that is, high confidence score indicates high accuracy and vise versa. In our experiments, data selection is performed on the utterance level so that the entire utterance is kept or rejected depending on its confidence score.

To apply the new data selection approach to semi-supervised acoustic model training, an obstacle has to be addressed: how to do clustering in a training set composed of utterances. For continuous speech, the length of utterance in terms of frame, phoneme or word is flexible. Therefore, utterances need be converted to vectors with fixed size so that K-Means algorithm can be performed to partition the utterance space. Our solution is proposed as follows. Suppose the system has a phone set with L phonemes that $Q = \{q_1, q_2, ..., q_L\}$, and the utterance to be converted consists of T frames and N phonemes, where the phoneme can be obtained by analyzing its transcripts or decoding hypothesis.

Utterance vector conversion:

1. Perform K-Means algorithm in frame vector space, partitioning the space into M clusters $\{\varpi_1, \varpi_2, ..., \varpi_M\}$.
2. Convert the T-frame utterance $u = [x_1, x_2, ..., x_T]$ into a M-dimension vector $v_1 = [\gamma_1, \gamma_2, ..., \gamma_M]$ where
$$\gamma_m = number\ of\ x_t \in \varpi_m / T \quad (7)$$
3. Convert the N-phoneme utterance $u = [y_1, y_2, ..., y_N]$ into a L-dimension vector $v_2 = [\eta_1, \eta_2, ..., \eta_L]$ where
$$\eta_l = number\ of\ q_l \in u / L \quad (8)$$
4. Concatenate v_1 and v_2 to make a $L+M$ dimensional vector v representing the utterance u.

By converting utterances into $L+M$ dimensional vectors, the K-Means algorithm can be applied to partition the utterance space into clusters $\{D_1, D_2, ..., D_K\}$. In our experiments we no longer estimate $P(c|D_k)$ since the number of possible hypothesis for an utterance could be infinite. Instead, we include the phoneme information in the utterance vector with the expectation that the class distribution can be implicitly modeled. Correspondingly, the selection of un-transcribed utterance within a cluster D_k will only de-

pend on the confidence score. Please note utterances are still selected across the entire feature space with the help of $P(D_k)$.

Acoustic Model Trained from	Word Error Rate
Transcribed Data	47.31%
Transcribed + Un-transcribed Data	44.41%

Table 2 Training with or without un-transcribed data

The initial acoustic model is trained using the 10 transcribed meetings. All the un-transcribed speech data are decoded with the initial model, and then appended to the transcribed set, along with their hypotheses, to train a new acoustic model. Table 2 presents the word error rate of these two models.

We further compare the new data selection approach with traditional approaches that use only high confidence scores to select utterances. In each iteration of incremental training, 20% of the un-transcribed speech data, measured by the number of frames, are selected to augment the current training set. Figure 4 plots the word error rates of the two approaches.

Figure 4 shows that the new approach is consistently superior to traditional method. When the first 20% un-transcribed are added to the training, our approach reduced the word error rate to 44.36% from 47.31%, representing a relatively 6% gain in performance, and began to outperform the model trained using all the un-transcribed data. In contrast, traditional method only reduced the error rate to 46.89% with the same amount of data. This indicates that the most suitable un-transcribed data for model training cannot be identified if we ignore its distribution in feature space. Both approaches reach their best performance, 43.18% and 44.02% respectively, when using 60% of the un-transcribed data.

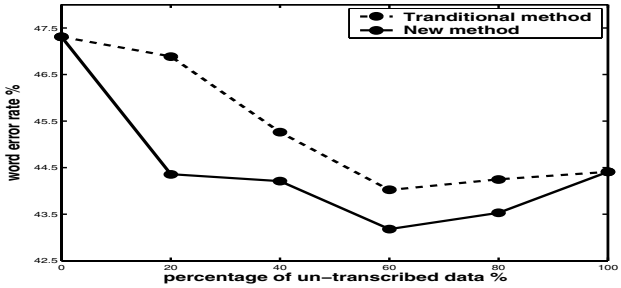


Figure 4 Comparison of two data selection methods with different amount of un-transcribed data

5. CONCLUSION

This paper investigates data selection methods used in semi-supervised incremental learning. Our analysis shows that traditional methods can lead to biased estimates, especially in the case where the confidence annotator is not independent of the classification model. We proposed a new data selection approach that in addition to using confidence score as the criterion to seek correctly classified examples, but also attempts to make the selection comply with true distribution. The effectiveness of the new approach is demonstrated by experimental results on a variety of machine learning and speech recognition problems.

6. REFERENCES

[1] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, Tom Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning, 39(2/3): 103-134 2000.

[2] Avrim Blum and Tom Mitchell, "Combining Labeled and Unlabeled Data with Co-Training", Proc. of the 11th Conference on Computational Learning Theory, 1998.

[3] Fabio G. Cozman, Ira Cohen and Marcelo C. Cirelo, "Semi-Supervised Learning of Mixture Models and Bayesian Networks", Proc. of 20th International conference on Machine Learning, 2003.

[4] Kamal Nigam and Rayid Ghani, "Analyzing the Effectiveness and Applicability of Co-Training", Proc. of the 9th International Conference on Information and Knowledge Management, 2000.

[5] P. J. Moreno and S. Agarwal, "An Experimental Study of EM-based Algorithms for Semi-Supervised Learning in Audio Classification", Proc. of ICML-2003 Workshop on Continuum from Labeled to Unlabeled Data, 2003.

[6] Thomas Kemp and Alex Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments", Proc. of 6th Eurospeech, 1999.

[7] Rong Zhang, Ziad Al Bawab, Arthur Chan, Ananlada Chotimongkol, David Huggins-Daines and Alexander I. Rudnicky, "Investigations on Ensemble Based Semi-Supervised Acoustic Model Training", Proc. of 9th Eurospeech, 2005.

[8] Frank Wessel, Klaus Macherey and Ralf Schluter, "Using Word Probabilities as Confidence Measures". Proc. ICASSP-98, Vol. 1, pp. 225-228, 1998.

[9] Yves Grandvalet and Yoshua Bengio, "Semi-Supervised Learning by Entropy Minimization", Proc. of 18th NIPS, 2004.

[10] Alicia Guerrero-Curieses and Jesus Cid-Sueiro, "An Entropy Minimization Principle for Semi-Supervised Terrain Classification", Proceedings of 7th IEEE ICIP, 2000.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelban, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolke, C. Wooters, "The ICSI Meeting Corpus" Proc. of ICASSP 2003, p. I-364, 2003.

Datasets	Labeling Rate	Training on Labeled Set	Semi-Supervised EM	High confidence Based Incremental Learning	New Data Selection Approach
		Acc. % (Mean ± Dev.)	Acc. % (Mean ± Dev.)	Acc. % (Mean ± Dev.)	Acc. % (Mean ± Dev.)
Image	5%	76.69±1.42	75.08±1.38	70.45±1.45	**77.86±1.25**
	10%	82.97±0.65	81.17±0.56	79.08±0.62	**83.13±0.58**
	20%	84.95±0.49	83.06±0.46	81.40±0.58	**85.53±0.50**
Letter	5%	63.95±0.92	64.08±0.98	67.19±0.97	**68.20±0.94**
	10%	73.77±0.80	74.85±0.88	77.38±0.88	**78.33±0.67**
	20%	78.01±0.73	79.23±0.79	82.64±0.65	**82.91±0.62**
Optdigits	5%	86.63±1.93	88.11±1.77	86.32±1.75	**88.51±1.54**
	10%	89.60±1.22	90.08±1.22	88.92±1.32	**90.56±1.05**
	20%	92.34±0.90	92.42±0.80	91.98±0.92	**93.31±0.81**

Table 1 Comparative study of four algorithms in multi-class classification