A New Data Selection Principle for Semi-Supervised Incremental Learning

Rong Zhang and Alexander I. Rudnicky
Language Technologies Institute, Carnegie Mellon University
{rongz, air}@cs.cmu.edu

Abstract

Current semi-supervised incremental learning approaches select unlabeled examples with predicted high confidence for model re-training. We show that for many applications this data selection strategy is not correct. This is because the confidence score is primarily a metric to measure the classification correctness on a particular example, rather than one to measure the example's contribution to the training of an improved model, especially in the case that the information used in the confidence annotator is correlated with that generated by the classifier. To address this problem, we propose a performance-driven principle for unlabeled data selection in which only the unlabeled examples that help to improve classification accuracy are selected for semi-supervised learning. Encouraging results are presented for a variety of public benchmark datasets.

1. Introduction

Semi-supervised learning has elicited growing interest in various research fields and many novel approaches have been proposed that give promising improvements in performance [1][2]. However some negative experiments showing that unlabeled data can deteriorate performance were also reported. The analyses of the possible reasons for the degradation point out that the increase of number of unlabeled examples can lead to a larger estimation bias when the model assumptions are violated [3][4]. This conclusion is quite discouraging since the situation it describes is very common in real-world applications (e.g., speech recognition, image and vision processing, information retrieval, etc.) for which the empirical models are only rough approximations to the true underlying distribution.

S_L: labeled set. S_U: unlabeled set.

Repeat until no example left in S_U:

- Train a model λ from S_L.
- Classify each example in S_U with λ
- V is the set of examples (with their classified label) picked from S_U according to some selection criterion.
- $S_L = S_L \cup V$.
- $S_U = S_U - V$.

Table 1: A version of semi-supervised incremental learning

Evidence suggests that semi-supervised incremental learning plus data selection can partially address the degradation problem [5][6][7][8]. Table 1 shows such an algorithm. Most semi-supervised incremental learning approaches share a notable characteristic: they select the unlabeled examples with the highest predicted confidence for model re-training. A first glance gives us the impression that this strategy is reasonable: high confidence scores usually imply that the class label assigned to an unlabeled example is correct, and expanding the training set with correctly labeled examples should be able to improve classification. However, there are counter-examples that question this strategy of data selection. For example, in continuous speech recognition, [9] reports that adding unlabeled data with lower confidence scores outperforms adding those with high confidence scores in improving recognition accuracy when combined with labeled data.

These contradictory observations suggest a more thorough investigation is needed. A brief analysis presented in the next section demonstrates that a confidence-based data selection strategy can lead to a poor estimate of the underlying distribution, especially in the case where the confidence annotator is constructed using information from the classifier. This paper introduces a solution to this problem: a novel performance-driven criterion that builds a bridge between data selection and classification accuracy. Specifically, with the new criterion, the selection of an unlabeled example for semi-supervised incremental learning is not based on whether its classification by the current model is correct, but rather on its potential contribution to the training of subsequent models. Confirming results on public benchmark datasets are presented.

2. Analysis on confidence-based data selection in semi-supervised learning

First, let's consider a question: is the confidence annotation model used in semi-supervised incremental learning independent of the classification model? One extreme example is co-training [5], in which the feature sets can be split into two independent and redundant subsets, each of which is sufficient for classification. Thus we can use one of them to construct a classifier, and the other to construct a confidence annotator for measuring the correctness of classification made by the former. However, many real-world applications cannot offer such a feature division that allows us to construct an independent confidence annotator. In these cases the confidence annotation is mainly based on the information supplied by the classification model. Therefore, the selections of unlabeled examples with high confidence score often result in only those examples that match well to the current model being picked. Re-training with such examples will consequently be a process that reinforces what the current model already encodes, yet it is unable to reduce the estimation bias caused by scarcity of labeled data or an inaccurate model assumption.

As we know, one advantage of semi-supervised learning is that the vast amount of unlabeled examples can help to generate an accurate estimate to the distribution of $P(x)$. This

raises the second question: do the unlabeled examples being selected for their high confidence score comply with the true distribution of $P(x)$? In our preliminary experiments with a variety of confidence metrics, we observed that the selected unlabeled examples often concentrate on some special regions of the input space, i.e. the region far from the class boundary, rather than distribute globally across the entire space. As a consequence, the $P(x)$ derived from the selected data will be an incorrect one. In addition, we also observed that in some applications there is no unlabeled example being selected for certain class, due to most of the examples of the class are located in a region classified with low confidence. Obviously, this will result in a biased estimate of the underlying prior probability of class $P(c)$. Although a biased estimate of $P(x)$ or $P(c)$ doesn't necessarily lead to the degradation of performance, it does make the model training more unpredictable especially when we use a generative model, i.e. Gaussian Mixtures, that need to learn $P(x,c)$ for classification.

Generally speaking, the confidence score is a metric that measures the classification correctness for a particular example, rather than a metric that evaluates the example's potential contribution for training an improved model. As suggested above, using high confidence score as the selection criterion may generate erroneous estimate of the true distribution, and thus may lead to the degradation of performance, even though the selected examples are correctly classified. We propose a performance-driven principle that attempts to address this problem from a new perspective: our solution no longer relies on a confidence metric; instead, the selection is made by evaluating the candidate's capability for improving classification performance. Specifically, only the unlabeled examples that help to increase classification accuracy are selected for model re-training.

1. Initialize each bin with empty set: $b_n = \Phi$ ($1 \le n \le N$).
2. Assign class label to each unlabeled example using current classifier λ, and compute confidence score for each classification using confidence metric $conf(c;x)$.
3. Split input space in to K subspaces $D_1, D_2, ..., D_K$ with reasonable clustering algorithm, i.e. K-Means.
4. For each subspace D_k ($1 \le k \le K$):
   - Sort the unlabeled example x that $x \in D_k$ according to its confidence score from high to low;
   - Add each unlabeled example x that $x \in D_k$ to one of the N bins in the fashion that b_n ($1 \le n \le N$) accepts the examples which confidence scores are within the range from top $\frac{n-1}{N}\%$ to $\frac{n}{N}\%$.

Table 2: Partition scheme

3. A performance driven principle for unlabeled data selection

The performance-driven principle is implemented as follows. The unlabeled set is first partitioned into a number of subsets (referred to as *bins* in this paper) as the candidates for data selection. An objective function is used to measure each bin's capability for improving classification accuracy. The bin of unlabeled examples achieving the best performance improvement, along with the automatically-assigned example labels, is added to the existing training set to train a new model. The process repeats until performance starts to deteriorate.

3.1. Partition scheme

Using the new principle, the unlabeled examples are partitioned into a number of equal-sized bins to create the candidates for selection. Two requirements are considered in our design of the partition scheme: the examples belonging to the same bin should contribute similarly to the model training; and they should be selected across the entire input space with respect to the distribution of $P(x)$. As the metric that reflects the degree to which an example matches the current model, we use confidence score to fulfill the first requirement. Meantime, a clustering algorithm, i.e. K-Means, is used to partition the input space into a number of clusters, so that the selection with $P(x)$ can be simulated by sampling examples from each cluster. Table 2 shows the detail of the algorithm.

3.2. Metric to evaluate model performance

Our goal is to identify those unlabeled examples that can improve system performance. This raises the question: how to measure the classification accuracy of a model in terms of labeled and unlabeled data. We use a metric that combines two aspects of information, *pseudo-accuracy* and *energy regularization*, in our experiments. Suppose there are *l* labeled examples (x_1, y_1), (x_2, y_2), …, (x_l, y_l), where class label $y_i \in \{c_1, c_2, ..., c_M\}$, and *u* unlabeled examples x_{l+1}, x_{l+2}, …, x_{l+u}.

Pseudo-accuracy. This item considers the classification accuracy of model λ on labeled and unlabeled sets. The calculation of accuracy on a labeled set is trivial. For unlabeled set, we assign a *pseudo-class*, which is the class label determined by majority voting among all the existing models obtained in iterations of semi-supervised incremental learning, to each unlabeled example. To allow more optimization methods i.e. *gradient descent* to be used for model training, *Pseudo-accuracy* is defined in the form of a continuous differentiable function.

We first define a discriminative function measuring model λ's capability to separate the desired class from other competing classes:

$$d_\lambda(x_i, y_i, \bar{y}_i) = \log P_\lambda(x_i, y_i) - \log([\tfrac{1}{M-1}\sum_{y \ne y_i} P_\lambda(x_i, y)^\eta]^{1/\eta}) \quad (1)$$

in which y_i is the *pseudo-class* when $l+1 \le i \le l+u$, and η is a parameter that controls how the competing classes are weighted. The interpretation of $d_\lambda(x_i, y_i, \bar{y}_i)$ is that, if it is negative, a classification error is assessed; otherwise, the classification is likely to be correct. $d_\lambda(x_i, y_i, \bar{y}_i)$ is then embedded into the *sigmoid* function, a smooth differentiable 0-1 function, to mimic the classification accuracy, and formulated as the metric *pseudo-accuracy*:

$$f_{acc}(\lambda) = \sum_{i=1}^{l+u} \frac{1}{1+\exp(-\rho * d_\lambda(\mathbf{x}_i, y_i, \bar{y}_i) + \theta)} \quad (2)$$

where parameters ρ and θ control the slope and transition point of the *sigmoid* function.

Energy regularization. Since the *pseudo-class* may be different from the correct class label, we regularize $f_{acc}(\lambda)$ using other metrics. Motivated by the observation that examples that are nearby in the input space often have the same class label, we add an energy based regularizer to the performance evaluation, which is defined as follows [10]:

$$f_{eng}(\lambda) = \sum_{i,j=1}^{l+u} w_{i,j} \sum_{m=1}^{M} [P_\lambda(c_m \mid x_i) - P_\lambda(c_m \mid x_j)]^2 \quad (3)$$

where $w_{i,j}$ is the weight of the edge linking example x_i and x_j, which value is inverse to their Euclidean distance.

$$w_{i,j} = \exp(-\|x_i - x_j\|^2) \quad (4)$$

In our implementation, x_j is restricted to the nearest neighbors of x_i.

The overall metric for evaluating model λ's performance is the linear combination of these two regularization factors.

$$f(\lambda) = \alpha_1 f_{acc}(\lambda) - \alpha_2 f_{eng}(\lambda) \quad (5)$$

where α_1 and α_2 are the weights for each regularizer, and set to 0.75 and 0.25 respectively in our experiments. Please note that in addition to being used in incremental learning, (5) can also be applied to general semi-supervised learning as a differentiable objective function allowing the use of method i.e. *gradient descent* for model optimization.

3.3. Performance-driven incremental learning

The performance-driven incremental learning algorithm is illustrated in Table 3. It differs from the traditional incremental learning algorithm described in Table 1 in that the selection of unlabeled data is based on their contribution to the improvement of classification. Specifically, the reason for bin b^* being picked for the next iteration is based on the classification accuracy that it helps to realize, which is measured by objective function $f(\lambda)$, is the best one among other bins, and more important, outperforms the accuracy of the current model. Moreover, early-stop criterion is adopted in the algorithm so that the training process can stop before the performance begins to degrade. Traditional approaches can be understood as special cases of this new algorithm, in which confidence metric is used as both data partition and selection criterion.

4. Experiments

To test the effectiveness of the proposed data selection approach for semi-supervised incremental learning, we performed a series of experiments on 7 benchmark datasets taken from the UCI machine learning repository: credit-screening (*crx*), glass-identification (*glass*), image-segmentation (*image*), iris-plant (*iris*), ionosphere, letter-recognition (*letter*), and optical-recognition-of-handwritten-digits (*optdigits*). These datasets span decision making, image recognition and speech recognition. For the two largest datasets, *letter* and *optdigits*, we adopt the pre-defined training/test split provided in their definition files: 16000/4000 for *letter* and 3823/1797 for *optdigits*. For the other 5 datasets, we divide the entire dataset into 10 equally-sized subsets, sequentially choosing one of them as the test set and the remaining as the training set (according to a cross-validation, or jack-knifing, procedure).

The labeled data are further separated from the training set by randomly selecting a certain portion of examples along with their labels. Two labeling selection rates are used in our investigations, 10% and 20%. To eliminate the uncertainty caused by random selection, the experiment is repeated for 200 times for each labeling rate and training/test split. The overall means and standard deviation of test accuracy are reported as the final performance. The number of bins is set to 10 which means 10% unlabeled examples are added to the existing training set in each iteration.

S_L: labeled set. S_U: unlabeled set.

Initialize:

1. Train the initial model λ_0 from labeled set S_L. Let $f_0 = f(\lambda_0)$ where $f(.)$ is the metric measuring model performance as shown in (5).
2. Partition the unlabeled set into *N* equally-sized bins $B_0 = \{b_1, b_2, ..., b_N\}$ using the scheme illustrated in Table 2.
3. Let $S_0 = S_L$ and $k = 0$.

Repeat if $B_k \neq \phi$:

1. For each bin $b \in B_k$,
   - Train a model $\lambda_{k+1,b}$ from $S_k \cup b$, where the class label of unlabeled example $x \in b$ is assigned by the current model λ_k.
   - Let $f_{k+1,b} = f(\lambda_{k+1,b})$.
2. Let $f_{k+1} = \max_{b \in B_k} \{f_{k+1,b}\}$.
3. If $f_{k+1} \leq f_k$, break and return λ_k as the training result.
4. Else, let $b^* = \arg\max_{b \in B_k} \{f_{k+1,b}\}$.
   - $S_{k+1} = S_k \cup b^*$.
   - $\lambda_{k+1} = \lambda_{k+1,b^*}$.
   - $B_{k+1} = B_k - b^*$.
   - $k = k + 1$.

Table 3: Performance driven incremental learning algorithm

Our experiments use Gaussian Mixtures Model (GMM) as the base classifier, K-Means as the clustering method in partition scheme and *Negative Entropy* [11][12] as the confidence metric defined as follows.

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

Table 4 presents the results of four learning approaches: (1) supervised learning on the labeled data, (2) semi-supervised learning using generalized EM [2], (3) traditional incremental learning that always selects high confidence data (as illustrated in Table 1), and (4) our proposed performance-

driven incremental learning algorithm. The best result for each dataset and labeling rate is marked in boldface type.

We first compare the performances of supervised learning and semi-supervised learning with EM. Degradations caused by using unlabeled data are observed in some of the datasets. These results remind us that regardless of the promising perspective of semi-supervised learning, additional investigations need to be done to understand the necessary and sufficient conditions for such degradations.

Table 4 shows that incremental learning appears to be an alternative option to generalized EM. However, the traditional incremental learning algorithm based on the selection of high confidence data is far from satisfactory, especially when no early-stop criterion is used. In contrast, our performance-driven incremental learning algorithm works very well, and consistently performs as well as or better than the best of the other three approaches. Decent improvements of classification accuracy are observed in most of the 7 datasets.

5. Conclusion

We investigated the data selection criteria used in semi-supervised incremental learning. We empirically demonstrate that the traditional criterion, focusing on unlabeled examples with high classification confidence is not necessarily the best choice. We introduce a novel performance-driven principle for unlabeled data selection: An unlabeled example is selected for inclusion based on its capability to improve classification accuracy. The effectiveness of this principle is demonstrated in a series of experiments on UCI benchmark datasets.

6. References

[1] Matthias Seeger, "Learning with labeled and unlabeled data", Technique Report, University of Edinburgh, 2002.

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

[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] Grigoris Karakoulas and Ruslan Salakhutdinov, "Semi-Supervised Mixture-of-Experts Classification", Proc. of 4th IEEE International Conference on Data Mining, 2004.

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

[6] 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.

[7] 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.

[8] Chunk Rosenberg, Martial Hebert and Henry Schneiderman, "Semi-Supervised Self-Training of Object Detection Models", Proc. of 7th IEEE Workshop on Applications of Computer Vision, 2005.

[9] 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 European Conference on Speech Communication and Technology, 2005.

[10] Xiaojin Zhu, Zoubin Ghahramani and John Lefferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions", Proc. of 20th International Conference on Machine Learning, 2003.

[11] Alicia Guerrero-Curieses and Jesus Cid-Sueiro, "An Entropy Minimization Principle for Semi-Supervised Terrain Classification", Proc. of 7th IEEE International Conference on Image Processing, 2000.

[12] Yves Grandvalet and Yoshua Bengio, "Semi-Supervised Learning by Entropy Minimization", Proc. of 18th Neural Information Processing System Conference, 2004.

Datasets	Labeling rate (%)	Training on labeled set	Semi-supervised EM	High conf. based inc. learning	Performance-driven incremental learning
		Accuracy Mean ± Dev (%)	Accuracy Mean ± Dev (%)	Accuracy Mean ± Dev (%)	Accuracy Mean ± Dev (%)
Crx	10	82.26±1.12	82.15±0.81	81.55±1.20	**82.69±0.93**
Crx	20	83.23±0.73	82.59±0.67	83.17±0.72	**83.54±0.71**
Glass	10	43.21±3.62	40.09±3.86	43.56±3.33	**44.52±3.49**
Glass	20	44.67±3.15	40.82±2.63	44.87±3.17	**46.49±3.22**
Image	10	82.95±0.76	81.29±0.74	83.23±0.66	**84.01±0.62**
Image	20	85.00±0.49	82.94±0.47	85.35±0.54	**85.80±0.51**
Iris	10	91.42±1.47	90.27±1.24	91.59±1.47	**92.61±1.53**
Iris	20	91.94±1.32	90.80±1.17	91.99±1.23	**92.87±1.35**
ionosphere	10	78.74±2.83	80.36±3.07	82.44±2.59	**83.59±2.52**
ionosphere	20	79.32±2.29	81.38±2.61	85.86±2.07	**86.93±2.04**
Letter	10	73.77±0.80	74.85±0.88	77.38±0.88	**78.33±0.67**
Letter	20	78.01±0.73	79.23±0.79	82.64±0.65	**82.91±0.62**
optdigits	10	89.55±1.37	90.12±1.33	89.83±1.67	**90.95±0.87**
optdigits	20	92.29±0.63	92.60±0.59	93.15±0.82	**93.50±0.61**

Table 4: Comparative study of four algorithms