A Co-training based Framework for Writer Identification in Offline Handwriting

Utkarsh Porwal
Deptt. of Computer Science and Engg.
University at Buffalo - SUNY
Amherst, NY - 14228
utkarshp@buffalo.edu

Venu Govindaraju
Deptt. of Computer Science and Engg.
University at Buffalo - SUNY
Amherst, NY - 14228
govind@buffalo.edu

***Abstract*—Traditional forensic document analysis methods have focused on feature-classification paradigm where a machine learning based classifier is used to learn discrimination among multiple writers. However, usage of such techniques is restricted to availability of a large labeled dataset which is not always feasible. In this paper, we propose a Co-training based approach that overcomes this limitation by exploiting independence between multiple views (features) of data. Two learners are initially trained on different views of a smaller labeled training data and their initial hypothesis is used to predict labels on larger unlabeled dataset. Confident predictions from each learner are used to add such data points back to the training data with predicted label as the ground truth label, thereby effectively increasing the size of labeled dataset and improving the overall classification performance. We conduct experiments on publicly available IAM dataset and illustrate the efficacy of proposed approach.**

***Keywords*-Writer Identification, Co-training, Classifier, Views, Labeled and Unlabeled data**

I. Introduction

Writer Identification is a well studied problem in forensic document analysis where the goal is to correctly label the writer of an unknown handwriting sample. Existing research in this area has sought to address this problem using Machine Learning techniques, where a large labeled dataset is used to learn a model (supervised learning) that efficiently discriminates between various different writer classes. The key advantage of such learning approaches is their ability to generalize well over unknown test data distributions. However, such generalization provides greater performance only when used with a large labeled data. In real-world scenarios, generating large labeled datasets requires manual annotation which is not always practical. The absence of such datasets also leads to inefficient usage of available unlabeled data that can be exploited to provide a greater classification performance. To address these issues, we propose a Co-training based learning framework that learns multiple classifiers on different views (features) of smaller labeled data and uses them to predict labels for unlabeled dataset which are further bootstrapped to the labeled data for enhancing the prediction performance.

Existing literature on writer identification can be broadly classified into two categories. First category is of text dependent features which capture properties of writer based on the text written. In this writer identification is done by modeling similar content written by different writers. This reliance on text dependent features poses challenges of scalability. In real world application such data is seldom available which limits the usability of these techniques for practical purposes. said et al. [14] extracted text dependent features using Gabor filters but the main limitation was to have a full page of document written by different writers for identification. Second category is based on text independent features. They capture writer specific properties such as slant and loops which are independent of any text written. These techniques are better suited for real life scenarios as they directly model writers as opposed to previous category. Feature selection plays an important role in such techniques. Several features capturing different aspects of handwriting has been tried. zois et al. [15] used morphological features and needed only single word for identification and niels et al. [17] used allographic features to compare using Dynamic Time Warping(DTW). All of this work was focused on better feature selection which would result in better accuracy. They did not lay stress on the techniques used and made an assumption that sufficient amount of such data is available for the system to learn

Likewise, writer identification can also be divided under two major approaches. First is statistical analysis of several features such as edge hinge distribution. Edge hinge distribution captures the change in the direction of writing samples. Second approach is model based writer identification. In this predefined models of strokes of handwriting are used. Prime focus of these techniques was on making a better system for identification using different techniques for modeling and analysis. Various techniques such as Latent Dirichlet Allocation(LDA) were proposed for higher accuracy for identification[12] but it was based on the assumption that sufficient training data is available.

Existing techniques and methods did not make use of unlabeled data for the identification. Information tapped in the unlabeled data can make a significant improvement in

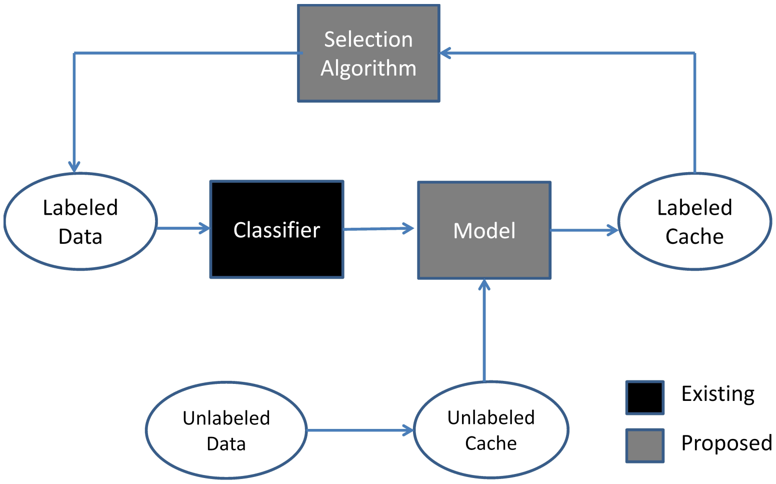

Figure 1. Schematic of Proposed Co-training Based Labeling Approach

the performance of the system. This information can be extracted using different techniques such as transductive SVMs[11] or graph based methods using EM algorithm. They are used to label unlabeled data in a semi supervised framework. nigam et al. [7] later proved that Co-training performs better than these methods in semi supervised framework. It uses small snippet of labeled data and iteratively labels some part of unlabeled data. System retrains itself after every iteration which results in better accuracy. Co-training has been successfully used for semi supervised learning in different areas but never been used for labeling data for writer identification to the best of our knowledge. Co-training has been used for web page classification[1], object detection[5] and for visual trackers[4] . It has been used extensively in NLP for tasks like named entity recognition[6].

The organization of the paper is as follows. Section 2 provides an overview of Co-training based framework. Multiple data views in form of writer features are described in Section 3. Section 4 illustrates the proposed approach. Experimental results are described in Section 5. Section 6 outlines the conclusion.

II. Co-training

Co-training is a semi supervised learning algorithm which needs small amount of training data to start. It reiteratively labels some unlabeled data points and again learns from it. blum et al. [1] proposed co-training to classify web pages on the internet into faculty web pages and non-faculty web pages. Initially they used small amount of web pages of faculty members to train a classifier and were able to correctly classify most of the unlabeled pages correctly in the end. Co-training requires two separate views of the data and two learners. blum et al. [1] proved that co-training works best if the two views are orthogonal to each other and each of them is capable of classification independently. They showed that if the two views are conditionally independent then the accuracy of classifiers can be increased significantly. This is because system is using more information to classify data points. Since both views are sufficient for classification, this brings redundancy which in turns gives more information. nigam et al. [8] later proved that completely independent views are not required for co-training. It works well even if two views are not completely uncorrelated.

Co-training is an iterative bootstrapping method which increases the confidence of the learner in each round. It boosts the confidence of score like Expectation Maximization method but it works better than EM[7]. In EM all the data points are labeled in each round while in Co-training few of the data points are labeled each round and then classifiers are retrained. This helps building a better learner in each iteration which in take would make better decision and hence the overall accuracy of system will increase.

A. *Selection Algorithm*

Selection of data points is crucial in the performance of the algorithm. New points added in each round should make learner more confident in making decisions about the labels. Hence, several selection algorithms have been tried to make a better system as system's performance can vary if selection method is changed. Different methods out performs each other depending on the kind of data and application. One approach to select points was based on performance[2]. In this method, some points were selected randomly and added to the labeled set. System was retrained and its performance was tested on the unlabeled data. This process was repeated for some iterations and performance of every set of points was recorded. Set of points resulting in best performance were selected to be added in the labeled set and rest were discarded. This method was based on the degree of agreement of both learners over unlabeled data in each round.

Some other methods has also been employed like choosing the top *k* elements from the newly labeled cache. This is an intuitive approach as those points were labeled with the highest confidence by the learner. However, hwa et al. [9] in their work showed that adding samples with best confidence score not necessarily results in better performance of classifiers. So, wang et al. [10] used a different approach in which some data points with lowest scores were also added along with the data points with highest confidence scores. This method was called *max-t, min-s* method and *t* and *s* were optimized for the best performance. So, several different selection methods have been employed as selecting data point in each round is key to the performance of Co-training.

III. Feature Selection

Selection of uncorrelated views is important in the working of Co-training. blum et al. [1] proposed that both views should be sufficient for classification. Each learner trained on the views should be a low error classifier. They proved that error rates of both the classifiers decreases during Co-training because of the extra information added to the system. This extra information directly depends on the degree of uncorrelation. However, abney et al. [3] later reformulated the explanation given by [1] for the working of Co-training in terms of measure of agreement between learners over unlabeled data. abney et al. [3] gave an upper bound on the error rates of learners based on the measure of their disagreement. Hence, independence of both views is crucial for the performance of the system. We chose contour angle features[13] as a first view and we combined structural and concavity features (SC)[18] as a second view. These features can be considered independent as both captures different properties of style of writing.

IV. Proposed Method

Co-training fits naturally for the task of writer identification as any piece of writing can have different views. Contour angle features and structural and concavity features are two such different views for any handwritten text. They can be considered uncorrelated enough to fit the task of writer identification in Co-training framework. Co-training also needs to have two learners to learn over two views. We used two different instances of Random Forest as learners to normalize the effect of learner over views.

Angle features were used to train first classifier and SC were used to train the other one. Then in each round a cache will be extracted from unlabeled data. This cache would be labeled by both learners and some data points will be picked from newly labeled cache by selection algorithm. Selected data points will be added to the training set and the learners are retrained while remaining data points in the cache are discarded. This process is repeated unless the unlabeled set is empty. Below is the pseudo code for the Co-training algorithm.

```
Algorithm 1 $Co-trainingAlgo$
Require:
  $L1 \leftarrow$ Labeled View One
  $L2 \leftarrow$ Labeled View Two
  $U \leftarrow$ Unlabeled Data
  $H1 \leftarrow$ First Classifier
  $H2 \leftarrow$ Second Classifier
  Train $H1$ with $L1$
  Train $H2$ with $L2$
  repeat
      Extract cache $C$ from $U$
      $U \leftarrow U - C$
      Label $C$ using $H1$ and $H2$
      $d \leftarrow$ selection_algo($C$) where $d \subset C$
      add_labels($d$,$H1$,$H2$)
      $L1 \leftarrow L1 \cup$ view one of $d$
      $L2 \leftarrow L2 \cup$ view two of $d$
      Retrain $H1$ on $L1$
      Retrain $H2$ on $L2$
  until $U$ is empty
```

A. Selection Algorithm

Selection algorithm used for selecting data points was based on agreement of both learners over data points. Points on which the confidence of both learners was above certain threshold were selected. In case of documents accuracy of classifier would be high if two different views will indicate same label for any data point. Selection method based on randomly selecting data points and checking their performance as used in [2] was not good as randomly checking takes time. The approach is not scalable as there are several rounds of processing of subset of cache every time a new cache is retrieved. Below is the pseudo code for the selection algorithm. Score function in the algorithm gives the highest value of the confidence scores of the learner for one data point over all writers.

Table I
ACCURACY OF CLASSISIERS WITH BASELINE SYSTEM AND CO-TRAINING

Methods	**Full Data**	**Half Data**	**One Fourth Data**	**One Tenth Data**
Experiment 1 Baseline	83.73	79.64	74.48	59.00
Co-training	85.58	80.91	75.55	61.24
Experiment 2 Baseline	80.42	76.72	70.59	52.28
Co-training	82.47	77.31	72.15	53.94

Algorithm 2 $SelectionAlgo$

```
Require:
  $C \leftarrow$ cache
  $t \leftarrow$ threshold
  $d \leftarrow \Phi$
  for each data point $c$ in $C$ do
      if score($c$,$H1$) $> t$ & score($c$,$H1$) $> t$ then
          $d \leftarrow d \cup c$
          $C \leftarrow C$ - $c$
      else
          $C \leftarrow C$ - $c$
      end if
  end for
  return $d$
```

V. EXPERIMENTS

We used IAM dataset which has total of 4075 line images written by 93 unique writers. We conducted two experiments to test the performance of Co-training against the baseline systems. In first we compared the accuracy of classifiers after Co-training against baseline methods by adding the scores of both learners. In this scores of the class distribution of the two learners were added for each data point and a joint class distribution score was generated. Class label with the highest score was assigned to that data point. Second experiment was based on the maximum of the confidence score of the label assigned by each learner. In this each classifier assigns a class label to the data point. This assignment is based on the highest value of the confidence score distribution over all classes. Class label with the higher score between the two is assigned to the data point.

Our goal is to show that Co-training can be used to label unlabeled data even if a small amount of labeled data is present in the beginning. Therefore experiments were run on dataset of different sizes. We conducted experiments with four different settings of data. System was initially trained over full, half, one fourth and one tenth of the total training data. In one tenth training data only three samples per class were present. Table shows that after Co-training accuracy of classifiers is better than the baseline system with all sizes of datasets in both experimental settings.

VI. CONCLUSION

In this paper we presented a Co-training based framework for labeling a large dataset of unlabeled document with the correct writer identities. Previous work in writer identification was focused on either on developing a better feature selection algorithm or to use different techniques for modeling the text of the document. All the work was based on a assumption that sufficient amount of labeled data is available for training a system. In our work we address the problem of limited amount of labeled data present in real life applications. Our method tries to iteratively generate more labeled data from unlabeled data. Experimental studies show that accuracy of learners on the dataset labeled by Co-training was better than the baseline system. This proves the effectiveness of Co-training for labeling a large dataset of unlabeled documents. In future we would like to address this problem of limited data by using other semi supervised learning methods.

REFERENCES

[1] A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, In Proceedings ofCOLT '98, pp. 92-100.1998.

[2] S. Clark, J. Curran, and M. Osborne, *Bootstrapping POS taggers using unlabelled data*, In Proceedings of CoNLL, Edmonton, Canada, pp. 4955. 2003.

[3] S. Abney, *Bootstrapping*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002

[4] O. Javed, S. Ali and M. Shah, *Online detection and classification of moving objects using progressively improving detectors*, In Computer Vision and Pattern Recognition, pp. 696-701. 2005.

[5] A. Levin, P. Viola and Y. Freund, *Unsupervised improvement of visual detectors using cotraining*, Proceedings of the Ninth IEEE International Conference on Computer Vision,ICCV '03

[6] M. Collins and Y. Singer, *Unsupervised Models for Named Entity Classification*, Empirical Methods in Natural Language Processing - EMNLP. 1999

[7] K. Nigam and R. Ghani, *Understanding the Behavior of Co-training*, In Proceedings of KDD Workshop on Text Mining, 2000.

[8] K. Nigam and R. Ghani , *Analyzing the effectiveness and applicability of co-training*, Proceedings of the Ninth International Conference on Information and Knowledge Management, pp. 86-93. 2000

[9] R. Hwa, *Sample selection for statistical grammar induction*, In Proceedings of Joing SIGDAT Conference on EMNLP and VLC, Hongkong, China, pp. 4552. 2000

[10] W. Wang, Z. Huang and M. Harper, *Semi-Supervised Learning for Part-of-Speech Tagging of Mandarin Transcribed Speech*, In IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.

[11] T. Joachims, *Transductive Inference for Text Classification using Support Vector Machines*.In Proceedings of the Sixteenth International Conference on Machine Learning. pp. 200-209. 1999.

[12] A. Bhardwaj, M. Reddy, S. Setlur, V. Govindaraju and S. Ramachandrula, *Latent Dirichlet allocation based writer identification in offline handwriting*In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. pp. 357-362, 2010

[13] M. Bulacu and L. Schomaker, *Text-Independent Writer Identification and Verification Using Textural and Allographic Features*, In IEEE Transactions on Pattern Analysis and Machine Intelligence. pp 701-717. 2007

[14] H. E. S. Said, G. S. Peake, T. N. Tan and K. D. Baker, *Personal identification based on handwriting*. Pattern Recognition, 33, pp. 149-160. 2000

[15] E. N. Zois and V. Anastassopoulos, *Morphological waveform coding for writer indentification*. Pattern Recognition, 33(3), pp. 385-398. 2000

[16] L. Schomaker and M. Bulacu, *Automatic writer identification using connected-component contours and edge-based features of uppercase Western script*. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 787-798. 2004

[17] R. Niels, L. Vuurpijl and L. Schomaker, *Introducing TRIGRAPH - Trimodal writer identification*. In Proceedings of European Network of Forensic Handwriting Experts, 2005

[18] J.T. Favata, G. Srikantan, S.N. Srihari, *Handprinted character/digit recognition using a multiple feature/resolution philosophy*, In Proceedings of Fourth International Workshop Frontiers of Handwriting Recognition. 1994.