Structural Learning for Writer Identification in Offline Handwriting

Utkarsh Porwal, Chetan Ramaiah, Arti Shivram and Venu Govindaraju
Department of Computer Science and Engineering
University at Buffalo - SUNY
Amherst, NY - 14228
utkarshp,chetanra,ashivram,govind@buffalo.edu

Abstract

Availability of sufficient labeled data is key to the performance of any learning algorithm. However, in document analysis obtaining the large amount of labeled data is difficult. Scarcity of labeled samples is often a main bottleneck in the performance of algorithms for document analysis. However, unlabeled data samples are present in abundance. We propose a semi supervised framework for writer identification for offline handwritten documents that leverages the information hidden in the unlabeled samples. The task of writer identification is a complex one and our framework tries to model the nuances of handwriting with the use of structural learning. This framework models the complexity of learning problem by selecting the best hypotheses space by breaking the main task into several sub tasks. All the hypotheses spaces pertaining to the sub tasks will be used for the best model selection by retrieving a common optimal sub structure that has high correspondence with all of the candidate hypotheses spaces. We have used publically available IAM data set to show the efficacy of our method.

1. Introduction

Writer identification has become an active area of research in the field of documents analysis where goal is to correctly identify the writer of any handwritten sample from a list of writers known in advance. It is an important task primarily because of the multitude of applications such as digital libraries, forensic document analysis and in smart devices like tablets and phones. Accuracy of the identification systems in the above mentioned applications is sought to be close to human accuracy but current technology is not even close and hence there is a large room for improvement in the performance of the systems.

State-of-the-art techniques for writer identification focuses either on exploiting the nuances of the text such as style, strokes etc or on learning a machine learning algorithm which once trained will generalize well on any new test sample. Major hindrance in both of these approaches is the lack of sufficient training data. For any machine learning model to work it needs to be trained first and more training data leads to a better model. However, for training labeled samples are required and obtaining large amount of handwritten samples by different writers is not practical when application's reach is worldwide. On the other hand handwritten samples without the knowledge of their writers are easily available. Hence, in such scenario leveraging the information trapped in unlabeled samples could be very useful in enhancing the performance of the system. We propose a structural learning based approach where a target task is divided into several related auxiliary tasks. These auxiliary tasks are then solved and a common structure among all is retrieved which in turn is used to solve the actual target tasks. This approach also known as multi task learning in machine learning literature is very effective in semi supervised framework where labeled data is limited and the original problem is complex and rich enough to be broken down into sub problem and then each candidate sub problem will give information useful for solving the actual target problem.

The organization of the paper is as follows. Section 2 provides an overview of the related work done for writer identification for handwritten documents. Section 3 outlines the underlying principle of structural learning. Section 4 gives the motivation behind this work and illustrates the working of the structural learning algorithm in detail. Section 5 describes the features and classifiers used and have the experimental details. Section 6 outlines the conclusion.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.277

2. Related Work

State-of-the-art methods for writer identification can be broadly divided under two approaches. The first approach is text dependent in which features encapsulates the characteristics of the writer based on the similar text content written by different writers. This approach is not extensible as similar content written by all authors is seldom available and it would be difficult to get the same content written by different people. Said et al. [8] extracted text dependent features using Gabor filters however their method required a full page of written text by different writers for identification which limits its usability for practical purposes. The second approach is based on text independent features. This approach captures the writer specific properties such as slant and loops which are independent of any text written. These techniques are better suited for real world scenarios as they are scalable as they directly model writers as opposed to the text like the methods based on first approach do. Feature selection plays an important role in such techniques. Several features capturing different aspects of handwritten text have been tried. Zois et al. [9] used morphological features and needed only single word for identification whereas Niels et al. [10] used allographic features for comparison. Likewise, statistical analysis of several features has been done such as edge hinge distribution. Edge hinge distribution captures the change in the direction of writing samples. Another approach is model based writer identification, where predefined models of strokes of handwriting are used.

Existing techniques and methods did not make use of unlabeled data for the identification. Information stored in the unlabeled data can make a significant improvement in the performance of the system. To make use of such information couple of techniques have been proposed such as transductive SVMs[4] and co-training[7]. In our previous work[3] we tried to solve this problem in a semi supervised framework using co-training. Now, we propose a new semi supervised framework for the task of writer identification using the structural learning. Structural learning has been used in past for different applications such as part of speech tagging[2] in NLP and text categorization[1]. However, it has never been used before for the task of handwriting identification.

3. Structural Learning

The concept of structural learning has been formulated and studied in the related areas of machine learning[1][2] earlier. However, for the completeness of the description the framework has been explained in the sub sections below.

3.1. Supervised Learning

A learning problem can be formally defined as finding the mapping function between the input vector $\mathbf{x} \in \mathbf{X}$ and its output labels $y \in \mathbf{Y}$. In training phase finite number of data points $(\mathbf{x_i}, y_i)$ are provided under the assumption that they are all drawn from some unknown probabilty distribution Ω. The function that will take input vector $\mathbf{x}$ and gives output label y is called target concept. Any learning algorithm tries to learn this target concept and outputs a hypothesis h from $\mathbf{H}$ to approximate the target concept with the least possible error which can be defined as

$$err_{\Omega}(h) := Prob_{(\mathbf{x},y)\sim\Omega}[h(\mathbf{x}) \neq y] \tag{1}$$

Learning algorithm explores the hypotheses space by minimizing error over number of samples. Hence, more training data will lead to a better hypothesis. If hypotheses space is small then it is possible to explore it with less number of training samples. However, it is more likely that the actual target concept may not lie within the hypotheses space because of the small size. Error because of bad quality of hypotheses space is called *approximation error*. Likewise, if hypotheses space is large and rich then it would require large number of samples to explore it. Error because of limited number of samples is called *estimation error*. Therefore, the selection of right hypotheses space is a tradeoff between above mentioned two errors and is a key problem in machine learning. However, domain knowledge and context based assumptions can help in the selection of right model and can be useful in increasing the performance of the algorithm. Therefore, any algorithm will output the hypothesis which is consistent with most of the training samples and will minimize the error. Therefore, formally a learning algorithm will output

$$h^* = \underset{h\in\mathbf{H}}{\operatorname{argmin}} \{\text{err}(h)\} \tag{2}$$

3.2. Hypotheses Space Selection

In any real world application training data is limited as labeling data is expensive as it is time consuming and often requires human expertise. However, we can explore the hypotheses space only with the help of training samples to approximate the actual target function. Since, limited data points are available to explore

any candidate hypothesis space, selection of appropriate and rich space is central to the performance of the learner. Often learner fails to approximate the target function because it does not lie within the space learning algorithm is exploring. Hence, the central idea of structural learning is to select the most appropriate hypotheses space with the use of finite labeled data available.

The key concept of structural learning is to break the main task into several related tasks and then find a common low dimensional optimal structure which has high correspondence with every sub task. This structure is used to solve the main problem. The optimal structure would correspond to the scenario where the cumulative error of all the sub tasks will be minimized. This is very intuitive and holds true for any real life scenario. It is often desirable to break the task into small tasks as solving small tasks gives insight for the solution of actual problem.

The efficacy of structural learning lies in the fact that in almost all of the real world problems the hypothesis given by an algorithm is a smooth discriminant function. This function maps points in the data domain to the labels. The smoothness of this function is enforced by a good hypotheses space. If any two points are close in the domain space then mapping produced by discriminant function will also be close in the target space. Therefore, if one can find such discriminant functions then this implies good hypotheses space.

In structural learning we find several such functions that correspond to the structure of the underlying hypothesis space. If sub tasks are related we get information about context embedded in the optimal structure discovered. If sub tasks are not related then also structural parameter contains the smoothness information of the hypotheses space. Therefore, breaking the main task into sub tasks is helpful even though they are not related as the structure retrieved will still have the information about smoothness of the space.

Formally structural learning can be defined as collection of T sub tasks indexed by $t \in \{1....T\}$ and each sub task has n_t samples over some unknown distribution Ω_t. All the sub tasks has their respective candidate hypotheses spaces $\mathbf{H}_{\theta,t}$ indexed by the parameter θ which is the common to all the sub tasks and encapsulates all the information that is useful for solving the primary task. The new objective function is to minimize the joint empirical error

$$h^*_{\theta,t} = \underset{h \in \mathbf{H}_{\theta,t}}{\operatorname{argmin}} \sum_{i=1}^{n_t} L(h(\mathbf{x}_i^t), y_i^t) \tag{3}$$

where L is the loss function.

4. Proposed Method

4.1. Motivation

Writer Identification of a handwritten document is a difficult task because of the several challenges it offers. Handwriting of an individual captures several nuances of writer's personality and background. Some of these can be analyzed with different aspects of the handwriting such as size, loops, slants and continuity. However, all the information encapsulated in the handwriting of an individual is not only limited to such visible features. There are factors influencing the handwriting of an individual which are abstract such as the effect of the native language on the writing of the non native languages also known as accent of an individual [6]. Likewise there are styles under which one can broadly fit most of the writing styles. These features are not tangible and any learning algorithm needs to learn all these factors to effectively identify the writer of any handwritten document.

Since handwriting of an individual captures large amount of information the target function for the task will be a complicated function. To approximate the target function the hypotheses space should be rich and there should be enough training samples to explore this space. However, it is difficult to obtain large amount of handwritten documents from every writer. On the other hand handwritten documents where writers are not known are available in abundance. Therefore, a mechanism is needed to make use of the unlabeled data to improve the performance of the learning algorithm. If a learner can gain some insight into the nuances of handwriting in general by analyzing styles, loops and slants which are present in every handwritten sample using the unlabeled data then this information can be used in training phase of the learning procedure.

Structural learning offers a mechanism to break any task into auxiliary tasks and then learn a common optimal structure to all auxiliary tasks. This optimal structure captures information that is domain specific and is very useful in solving the main task as it helps in the selection of right hypotheses space. All the nuances of handwriting captured by accent, styles etc can be considered as related sub tasks of the main task of handwriting identification. Since as discussed above several such aspects of handwriting are abstract in nature there needs to be a principled way to define these sub tasks. Ando et al. [1] gave an approach to create such related sub tasks in a semi supervised framework to address the issue of limited training samples. In our work we will follow the same approach to create sub tasks with an assumption that they represent several different tangi-

ble and intangible aspects of handwriting.

4.2. Algorithm

The structural correspondence learning can be used to find out the common structure between different domains [2] or subtasks [1]. The idea is to seek the lower dimensional space that has high degree of correspondence with the primary domains or subtasks. In our work we propose the use of structural correspondence learning for distinct features of the same data. The first step is to create auxiliary tasks related to the main task. Auxiliary tasks can be sub tasks such as style identification and accent identification where main task is handwriting identification as discussed in previous section. Auxiliary tasks can also be formulated as capturing abstract aspects of the main tasks which cannot be formulated in real aspects such as styles and accents. However, they are vital in determining the writer of the handwritten document. Although there are no predefined methods to create auxiliary tasks but Ando et al. [1] suggested some generic ways to create auxiliary tasks.

One way to create auxiliary tasks in semi supervised framework is by making use of unlabeled data. However, creation of auxiliary task should address two issues. First is the label generation for the auxiliary tasks. The process should generate the automatic labels for each auxiliary task. Second condition is of relevancy among the auxiliary tasks. It is desirable that the auxiliary tasks are related to each other so that a common optimal structure can be retrieved. Ando et al. [1] suggested few generic methods to create auxiliary tasks that would satisfy these two conditions. In this work we followed one of those techniques.

In this approach two distinct features ϕ_1 and ϕ_2 are used. First a classifier is trained for the main task using the feature ϕ_1 over labeled data. Same feature is extracted from unlabeled data and the classifier trained is used to create auxiliary labels for the unlabeled data. The auxiliary task is to create binary classification problems for predicting the label assigned for each of the data point in unlabeled data. Therefore, for an *n* class problem as the main task *n* auxiliary tasks can be created as a two class problem. An auxiliary predictor will give label 1 if it can predict the correct auxiliary label otherwise it will assign 0. Any auxiliary predictor can be written as

$$h_{\mathbf{w}}(\mathbf{x}) = w_1 x_1 + w_2 x_2 + ... + w_n x_n \tag{4}$$

and goal is to reduce the empirical error as given by equation 2 and 3. Therefore, error can be written as

Algorithm 1 $Structural Learning Algorithm$

Require:

$X1 = [\phi_1(\mathbf{x})_t, y_t]_{t=1}^T \leftarrow$ Labeled Feature One
$X2 = [\phi_2(\mathbf{x})_t, y_t]_{t=1}^T \leftarrow$ Labeled Feature Two
$U = [\phi_2(\mathbf{x})_j] \leftarrow$ Unlabeled Data Feature Two
$C \leftarrow$ Classifier
Train C with $X1$
Generate auxillary labels by labeling U with C
For a L class probelm create L binary prediction problems as auxillary tasks, $y_l = h_l(\phi_2(\mathbf{x})), l = 1...L$
for $l = 1...L$ **do**
    $\mathbf{w}_{l,\theta} = (\phi_2(\mathbf{x})^T \phi_2(\mathbf{x}))^{-1} \phi_2(\mathbf{x})^T y_l$
end for
$\mathbf{W} = [\mathbf{w}_1|....|\mathbf{w}_L]$
$[\mathbf{U}\Sigma\mathbf{V}^T] = SVD(\mathbf{W})$
Projection onto $\mathbb{R}^h$, $\Theta = \mathbf{U}_{[:,1:h]} = [\theta_1|....|\theta_h]$
New feature in $\mathbb{R}^{N+h}$ space, $[\theta^T \phi_2(\mathbf{x}) \quad \phi_2(\mathbf{x})]$

$$y = h(\mathbf{w}, \mathbf{x}) + \epsilon \tag{5}$$

In order to minimize this error we take least squares loss function and minimize the joint empirical error for all the data points

$$err_\Omega(h) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x})^2 \tag{6}$$

To minimize the joint empirical error algorithm seeks the optimal weight vector. Setting gradient of the error function to zero will give the optimal weight vector

$$0 = \sum_{i=1}^{n} y_i \mathbf{x}^T - \mathbf{w}^T (\sum_{i=1}^{n} \mathbf{x}\mathbf{x}^T) \tag{7}$$

Solving for $\mathbf{w}$ we obtain

$$\mathbf{w}_{opt} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T y \tag{8}$$

This will give optimal weights for one predictor of one auxiliary task. To get the optimal structure corresponding to all the subtasks this process should be repeated for all the auxiliary tasks. After the optimal $\mathbf{x}$ is calculated for all the auxiliary tasks a big weight matrix $\mathbf{W}$ of all such weight vectors is created whose columns are the weight vectors of the hypothesis of auxiliary classes. In our work we have closely followed the work done by Blitzer et al. [2] and Ando et al. [1]. All the steps are described in Algorithm 1 shown above.

4.3. Dimensionality Reduction

Once the big weight matrix $\mathbf{W}$ is calculated it can be used to find the low dimensional common sub space. However, before doing dimensionality reduction redundancy in the information is removed. Often subtasks are related to each other along with the main task and they capture information of the same nature. Thus, it may not add much to discriminatory power of the algorithm to solve the main task. Since, information hidden in the weight vectors could be related only left singular vectors are picked from the singular value decomposition(SVD) of the $\mathbf{W}$ matrix. .Therefore,

$$[\mathbf{U}\Sigma\mathbf{V}^T] = SVD(\mathbf{W}) \tag{9}$$

Initially weight vectors are in the feature space $\mathbb{R}^N$ but they can projected onto some lower dimensional space $\mathbb{R}^h$ to capture the variance of auxiliary hypotheses space in best h dimension. Therefore the low dimensional feature mapping is $\theta^T\mathbf{x}$. In this work, we will append this feature mapping to the original feature vector and try to solve the main task in $\mathbb{R}^{N+h}$ space.

5. Experiments

5.1. Features

In this work, we have used GSC [11] and contour angle [5] features as two distinct features. Angle feature captures the orientation and curvature information of the handwritten characters which helps in characterizing the handwriting of an individual uniquely. We have used angle features for creating the auxiliary labels on the unlabeled data set. For creation of auxiliary tasks and getting a feature into a richer high dimensional space we have used GSC features.

GSC features has been known to capture such structural information very efficiently. They have been successfully used in several documents applications. They extract the local, intermediate and global information of the text. It takes a multi resolution approach by capturing the information at different levels as gradient features, structural features and concavity features. Gradient features provide local information about the stroke shape at shorter distance while structural features provide stroke shape information at longer distances. At global level concavity features captures relationship between different strokes.

5.2. Details

We have used IAM data set of 4075 line images from 93 different writers for conducting experiments. We have conducted experiments in four folds where each fold had around 2000 data samples as training data, around 1000 data samples as testing data and around 1000 data samples were considered as unlabeled to create auxiliary tasks and generation of weight matrix. Throughout the experiments we have used radial basis kernel SVM as our classifier. We used SVM for auxiliary label generation and later for the classification of the main task of writer identification. We have used 36 dimensional contour based angle feature for training an SVM for auxiliary label generation. For the generation of auxiliary tasks we have used 512 dimensional GSC feature. We extracted features by dividing the image into 4x4 frames where each frame will return 32 dimensional feature corresponding to first 12 as gradient feature, next 12 as structural features and the last 8 as concavity features. Therefore, from the full image we had 512 dimensional feature in which first 192 dimensions corresponds to the gradient feature, next 192 signifies structural features and last 128 are concavity features. All the features were scaled using z-score normalization. The baseline for the experiments is by using just the 512 dimensional feature vector for classification without adding any new features which correspond to the extra information encapsulated by the common optimal sub structure.

The main task in this work is a 93 class classification problem where writer of any handwritten sample is to be identified from one of the 93 writers. As discussed in previous sections the number of auxiliary tasks created for a c- class problem is c. Therefore, in this work we created 93 auxiliary tasks as 93 binary classification problems. Hence, the $\mathbf{W}$ matrix is of 512x93 dimension. We projected to a low h-dimensional space from original feature space to get the mapping in to that space. New mapped feature were appended to the 512 dimensional feature vector to make it $512 + h$ dimensional vector. This new feature vector of higher dimension is again normalized and an SVM was used for classification. The low dimensional optimal sub structure Θ, once retrieved, is used in both training and testing phase of the main task of writer identification. Results of experiments conducted with different h values are shown in table above. It can be observed through the experimental results that extra information provided by the common substructure is helpful in the identification task of writers given handwritten samples. Although selecting different low dimensional space did not result in much improvement as the number of auxiliary tasks

Table 1. Accuracy of our method at different dimensionality of optimal sub structure

Dimension of reduced space(h)	Fold 1	Fold 2	Fold 3	Fold 4	Average
Baseline	71.00	80.02	80.13	85.44	79.14
10	71.67	82.38	80.23	88.41	80.67
30	72.62	82.15	80.33	86.93	80.50
50	72.52	82.75	80.62	87.02	80.72
70	72.81	83.26	80.52	86.73	80.83
93	73.67	83.67	80.91	87.12	81.34

were already low. More the number of auxiliary tasks more information can be captured. Therefore, using the whole information resulted in the best performance.

6. Conclusion

In this paper we proposed an algorithm based on structural learning for the task of writer identification. The principle behind the working of the method is to break the target task into several related sub tasks as it is often difficult to model the complex structure of the target task. Many real world problems are complex in nature and can be divided into combination of sub tasks. It is intuitive and often easy to break down the problem and solve the individual problem and then solve the actual problem. In this work we showed that by breaking the task of writer identification and then retrieving the information shared by sub tasks in the form of a low dimensional common sub structure helps in improving the performance of the system for writer identification.

References

[1] R.K. Ando and T. Zhang, *A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data*, In The Journal of Machine Learning Research. volume 6. 2005

[2] J. Blitzer, R. McDonld and F. Pereira, *A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data*, In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2006

[3] U. Porwal, S. Rajan and V. Govindaraju, *An Oracle-based co-training framework for writer identification in offline handwriting*, In Proceedings of Document Recognition and Retrieval. 2012

[4] 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.

[5] 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

[6] C. Ramaiah ,U. Porwal and V. Govindaraju, *Accent Detection in Handwriting based on Writing Styles*, In Proceedings 10th IAPR International Workshop on Document Analysis Systems. 2012 (to appear)

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

[8] 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

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

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

[11] 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.