2012 10th IAPR International Workshop on Document Analysis Systems

Ensemble of Biased Learners for Offline Arabic Handwriting Recognition

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

***Abstract*—Techniques and performance of text recognition systems and software has shown great improvement in recent years. OCRs now can read any machine printed document with good accuracy. However, the advancements are primarily for Latin scripts and even for such scripts performance is limited in case of handwritten documents. Little work has been done for cursive scripts such as Arabic and still there is a room for improvement both in terms of accuracy and techniques. This paper presents an algorithm to recognize handwritten Arabic text using an ensemble of biased classifiers in a hierarchical setting. We address the fundamental shortcomings of the traditional Machine Learning paradigms when applied to Arabic scripts. Experiments have been conducted on the AMA Arabic dataset to show the efficacy of our method.**

***Keywords*-Ensemble, Arabic, Handwritten Text Recognition, Biased Classifiers**

I. Introduction

Automated offline handwriting recognition is an active area of research in document analysis because of the various challenges it offers. The goal here is to make a system that can recognize or read handwritten text with the same ease as humans do. However, there are several factors which inhibit us to achieve this goal and they will be discussed in great detail in the upcoming sections. Therefore there is still a scope for improvement in the performance of the recognition systems. Current OCR software can recognize the clean machine printed documents correctly but performs poorly for handwritten text inputs. OCR often fails to recognize a handwritten document of a slightly bad quality which is easily readable by humans. The primary reasons for failure are inconsistency in writing of a writer, different writing style of different writers for similar text and noise in the data. All these problems aggravate in case of complicated scripts such as Arabic. Heavy use of diacritical marks and the cursive nature of Arabic script are few such issues that need to be addressed . These problems make the recognition of even machine printed text a difficult task let alone handwritten texts. Therefore, analysis of Arabic scripts requires different feature extraction and learning techniques than traditional techniques employed for Latin scripts.

Machine learning techniques have been extensively used to solve the problem of handwriting recognition. A typical approach is to learn a classifier with sufficient writing samples and then use it to recognize the handwritten text in the future. Lot of factors play an important role in the performance of a classifier such as good feature selection and sufficient amount of training samples. Often, in real world applications learning one such classifier is difficult because of the reasons mentioned above. In such cases, ensemble of classifiers is a widely used technique which performs better than a single classifier. Idea is to learn multiple classifiers which will learn different subtasks and then in the end their collective opinion will be used for the primary classification task. There could be different ways of constructing a good ensemble of such classifiers. In this paper we propose an algorithm to perform classification using ensembles of biased classifiers in a hierarchical setting over an Arabic PAW dataset. Classifiers used here are biased towards specific classes and are good at recognizing one of the classes of PAWs with higher accuracy.

The organization of the paper is as follows. Section 2 provides an overview of the related work done for Arabic text classification. Section 3 outlines the causes of failure of learning algorithms and how ensemble is effective in such cases. Section 4 illustrates the nuances of Arabic scripts and highlights the reasons for the need of different approach for their analysis. Section 5 describes our proposed approach. Section 6 provides the details of the experimental setup and Section 7 outlines the conclusion.

II. Previous Work

One of the traditional approaches for handwritten text recognition is by segmenting the original text into smaller components. These components could either be characters or sub words. A learning algorithm is then applied on these smaller units which makes recognition of the words easy. However, segmentation based approaches may not be the best way in case of Arabic scripts because of several reasons. Firstly, Arabic scripts are always connected be it handwritten or machine printed. Hence, breaking the words into smaller segments results in new connected components which are never been learned and therefore makes recognition difficult. Secondly, any word in Arabic is not a single entity. It consists of dots and diacritics with different possible positions. Placement of these dots changes the context of the word so same set of graphemes can create different words. Therefore, segmentation of words is not a good approach as it may lose the contextual information of the structure

978-0-7695-4661-2 2012

DOI 10.1109/DAS.2012.35

as a whole. Segmentation has been done in past based on different parameters such as based on curvatures, strokes[1] and by dividing word into groups of letters[2]. AbdulKader et al.[10] used Neural Nets in a two-tier approach for offline Arabic handwriting recognition.

To overcome the shortcomings of segmentation based approaches Hidden Markov Models have been used successfully[3]. They perform better than the segmentation approaches as they capture the global structure by modeling the relation between all units of a word or sentence. However, HMMs have their own limitations as they assume the conditional independence of observations given hidden states. This is often not true as in case of Arabic scripts characters are connected because of the cursive nature. Therefore, this problem of recognition of Arabic scripts still remains an open problem with no standard approach. We propose an ensemble based method which addresses the fundamental limitations faced by traditional classifiers to learn the entire structure of the text.

III. Supervised Learning Analysis

Ensemble learning is a well formulated concept to learn a task with the help of multiple learners where a single learner would not suffice. Let's see why it works better than a single learner in certain cases. In a standard supervised learning problem an algorithm is provided with training data points in the form of

$$\{(\mathbf{x_1}, c(\mathbf{x_1})), (\mathbf{x_2}, c(\mathbf{x_2})) \ldots (\mathbf{x_n}, c(\mathbf{x_n}))\}$$

where $c(\mathbf{x})$ is the function that generates the labels of the corresponding data points $\mathbf{x}$. $c(\mathbf{x})$ is often known as the target function or target concept. Here $\mathbf{x}$ is a multi-dimensional vector in feature space where each individual element x_i represents some information about the data. It is assumed that all the training and test data points are sampled independently from the same distribution Θ which is the instance space. These are the basic assumption that has been made about the structure of the data that we are trying to learn. Now, any learning algorithm tries to learn a target concept c and it outputs a hypothesis h to approximate the target concept with the least possible error. Hypothesis generated by the learner should be consistent with the training data and should come from a predefined space of potential hypotheses called the hypothesis space χ. For any learning algorithm to perform well it is important that the target concept lies within that hypothesis space χ and the algorithm should be able to locate it.

Dietterich et al.[4] explained three primary reasons for learning algorithms to fail. Using an ensemble we will try to address these problems in this work. First reason stated is statistical in which learning algorithm searches the hypothesis space χ to find the target concept. If the amount of training data is not sufficient to search the entire space then algorithm outputs a hypothesis which fits the training data best. Since the training data is small there could be multiple such hypotheses and neither of them could be a good approximation to the actual target concept. In such cases algorithm fails to learn the actual parameters of the distribution Θ and performs poorly on the test data samples. Ensemble can be used effectively to address the issue of multiple consistent hypotheses. An ensemble can be constructed of multiple correct hypotheses and voting can be done to approximate the target concept. Second reason explained is computational in which learning algorithm fails to reach target concept because of computational reasons. Often algorithms perform local search to optimize cost functions and gets stuck in local minima's like gradient descent algorithms such as Artificial Neural Networks. Choice of a good starting point is essential for the good performance of such algorithms. Using ensemble of learners we can run different learners from different starting point to get a good approximation of the target concept. Third reason explained is representational as it is likely that the correct target concept cannot be represented by any of the hypotheses. It could be because learning algorithm stops searching the hypothesis space once it finds a good fit for the training samples. However by using ensemble of learners it is possible to get the better representation of the target concept by taking weighted sum of the individual hypothesis which will in turn expand the collective hypothesis space. Therefore, due to above mentioned arguments ensemble works better in cases of complex learning problems.

IV. Arabic Script

Arabic is a widely spoken language and it influences other languages such as Farsi and Urdu. Therefore, automated systems for recognition of Arabic text will have widespread benefits. However, the cursive nature and heavy use of supporting characters poses challenges to the state of the art techniques for Arabic text recognition. It is primarily because all the techniques have been developed for Latin scripts and cannot be applied directly to the Arabic scripts as two scripts are of different nature. Arabic has 28 letters and each letter can have more than one shape and selection of the shape depends on the position of the letter within its word or sub word. Position of the letter can be classified in four categories as in the beginning, middle and at the end of the word or it could be in isolation. Few letters also have *ascenders* and *descenders* which are written above or below the primary line. Other than the regular letters dots and diacritics are also used to represent vowels[5].

Hence letters in Arabic are highly dependent on the context in which they are written. Traditional approaches that use segmentation and modeling characters with HMMs are effective in learning the individual characters but fails to capture the contextual information embedded in the text. Since, this information is critical in the recognition of text we must have a learning algorithm to capture all

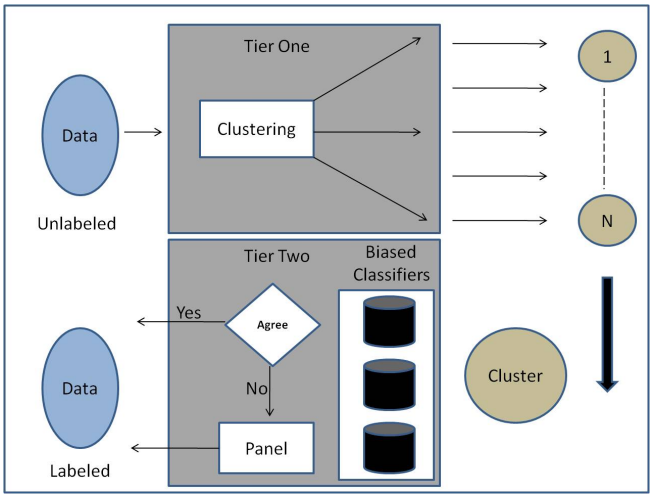

Figure 1. Schematic of Proposed Biased Learners Based Approach in a Hierarchical Framework

this information effectively. It is evident that the learning algorithm needs to learn a lot more factors to capture the structure of the entire text than just the structure of the characters. Therefore the hypothesis space for the learning algorithm must be very expressive. Since the distribution of the data is complex therefore hypothesis that learning algorithm outputs should also be complex. However, we have already discussed in previous section that in any real life application often there is not enough training data that one can effectively search the entire hypothesis space which is complex enough and can find a good approximation to the target concept. In such cases ensemble of learners comes to rescue for the reasons explained in the section above. Therefore, for complex learning tasks such as Arabic handwriting recognition ensemble of learners is a wise choice.

V. Proposed Approach

We propose an algorithm for classification of Arabic parts of words (PAWs) using biased classifiers in a hierarchical framework. As discussed in the section 4 Arabic scripts have a complex structure because of several reasons such as their cursive nature and contextual information hidden in the structure of the text. Therefore, it is important for any learning algorithm to have a hypothesis space rich enough to express all the complexities of the target concept. However, limited training data causes deterrence in this process and often just one kind of labeling is not enough to explore the structure efficiently. Hence we propose a novel way to classify with a hierarchical approach to reduce the complexity of the hypothesis space to be explored. We cluster the given training samples based on class labels to generate a new set of labels for the data. Intuition behind doing the clustering first is to group PAWs which are similar in some ways. Once this grouping is done a separate algorithm can be run to distinguish the member classes. This step helps in reducing the complexity of the task and a new label set will help in learning the actual structure of the data as it provides additional information. After clustering each data point has two types of labels one corresponding to the cluster and the other one is the actual PAW class it belongs to. A separate classifier will be learned with the new set of labels generated after clustering which will be unbiased. This is a first level of classification in the hierarchy. Efficacy of level one can be observed in figure 2 where accuracy of each cluster formed in first level of classification is plotted.

In level two we will use an ensemble of biased learners along with an arbiter panel. Our approach in level two is inspired from the work of Khoussainov et al.[6] in which focus was on individual classes instead of different regions of instance space to find the optimal discriminant function. This can be explained easily in case of a two class classification problem. Idea is to construct an ensemble to focus on individual classes by training a separate learner for positive and negative class. This can be done by constructing base learners such that they are biased towards individual classes. There should be a learner that is biased towards positive class by training it on data with majority of data points from positive classes so that it can easily identify positive class with higher accuracy in test data points. Likewise, a biased learner for negative class can also be learned by training a learner over data with class imbalance. In the proposed algorithm labels of data points where both learners will disagree are decided by an arbiter panel.

An arbiter panel is a group of classifier with different kind of inductive biases. Inductive bias is the *assumptions made by algorithm in addition to observed data to transform its output into logical deductions*[7]. There are several types of inductive biases such as maximum margin, nearest neighbor or maximum conditional independence. It plays an important

role when a learning algorithm outputs a hypothesis. If data points which are difficult to classify even by using biased classifier then it is intuitive to change the basic assumption made by the learner in calculating the discriminant function. Hence, in the arbiter panel learners with different inductive bias will be used to classify the data points with disagreement.

However Khoussainov et al.[6] followed a different approach to handle data points over which biased learners had disagreement. They trained a single arbiter classifier to deicide the label of the data point in case of disagreement. In their approach algorithm is run iteratively by re-weighting the data points in each round. Data points which remains misclassified after all iterations are said to be *hard* and are used as a training set for the arbiter classifier. Often data is not sufficient to run an algorithm several times or there could be a problem of class imbalance which prevents such approach as data of some classes are very few. Therefore, for this application we used an arbiter panel where all the classifiers are trained over the entire training data.

A. Multi Class Problem

Once idea of biased classifier is formulated for a two class problem it can be easily extended to multi class problems as well such as Arabic text classification. There are two standard ways of doing multi class classification. One is *one vs one* approach and other is *one vs all* approach. In *one vs one* approach a separate classifier is trained for each pair of classes as a two class problem and final classification is done by taking vote. In this approach number of classifier trained will be $\binom{n}{2}$ where n is the number of classes. For large number of n total number of classifiers will be large and to run such algorithm will be computationally expensive. Other way to do multi class classification is with *one vs all* approach. In this approach a two class classification problem is created for each class considering rest of the classes as the second class. In this approach total number of classifiers trained is equal to the total number of classes. In our work we have used *one vs all* approach.

B. Feature Selection

As discussed in sections above Arabic scripts have a complex structure where letters with different context differ by just few dots and diacritics. Therefore, feature extracted should be able to capture the entire structure of the text robustly. It should not only capture the local information about loops, slants and strokes but it should also be able to capture the global relationship that holds between the different components of the text. GSC features has been known to capture such structural information very efficiently[8]. They have been successfully used in several documents applications. GSC features extracts 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. Therefore GSC features fits well to the need of capturing structural information for the task of classification of Arabic PAWs efficiently.

VI. Experiments

We conducted experiments on the AMA dataset of Arabic part of words (PAW) which has 7312 images of 34 different types of PAWs. 6464 of the images are used for training and 848 images are used for testing the performance of the system. Distribution of data points corresponding to each of the PAW type is not uniform in the dataset. There is a class imbalance problem in this dataset as some classes have large number of data points and some have very few. This problem of class imbalance makes learning more difficult as learner could not learn the exact pattern of classes with insufficient data points. This can be observed in the figure 3 where class wise accuracy is plotted against the number of samples available (Numbers are scaled down to bring values in same range) for each class. Classes with relatively more samples were classified with better accuracy. The rise and drop in accuracy with the number of samples available can be easily observed.

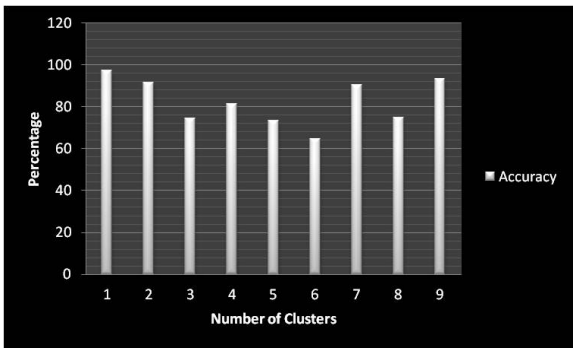

Figure 2. Accuracy of the System after Step One

We used two steps for classification of the test data samples. First, clustering was done on the test data samples and each data sample was assigned a cluster. We clustered the data based on the confusion matrix build in the training phase. Training data set was divided into train and validation set. We learned an SVM on the train data and it was tested over the validation set. A confusion matrix was build to see which classes can be put together. All the classes which were *confused* with each other were clustered into one class. Therefore, new labels were assigned to the data samples and a separate classifier was learned over the newly labeled

Table I
PERFORMANCE OF THE PROPOSED METHOD WITH BASELINE METHODS

Methods	Accuracy in Percentage
Accuracy of step 1	89.15
Accuracy of step 2	92.46
Overall Accuracy	82.42
Chen et al. Method	81.6
SVM	81.95

samples. In the first step data points were assigned cluster labels using the classifier learned over newly labeled samples and then step 2 is followed for each cluster.

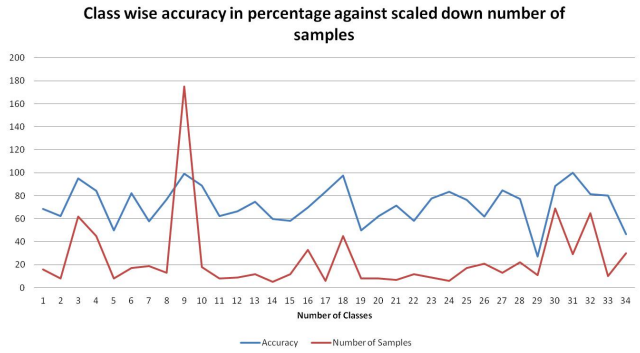

Figure 3. Overall Accuracy of the System

In step 2 numbers of classes has reduced to the classes in that cluster. Therefore, classifiers will be learned for the cluster members only. It is to be noted that upper bound on the performance of step 2 is the accuracy of the step 1. In this phase a separate learner is learned for each of the member class which is biased towards that class. Moreover, a panel of classifier is also learned for the cluster members who will act as a jury or arbiter panel in case of disagreement. We have chosen Naive Bayes, Random Forest and SVM for our arbiter panel because of their different kind of inductive bias. Therefore, any data point is first tested by the biased learners and if they disagree then arbiter panel will decide the label of the test sample. We have compared our result with the performance of Chen et al. [9] method using GSC feature. Table one show the performance over this dataset.

VII. CONCLUSION

In this paper we presented an algorithm to classify Arabic PAW dataset in a nontraditional way. We analyzed the fundamental shortcomings of the learning process and addressed the issues because of which a learner fails. We proposed a two step classification process which reduces the complexity of the instance space by assigning new labels to the data points. Redundancy in the information about the data helps in learning the structure of the data as exploring a new instance space of less complexity is easy. Once the area of interest could be located a biased learner approach can be employed to get the actual label of the data. Hence, this paper outlines a general method that can be applied to learning problems which are complex in nature because of various reasons such as intricate data structure or class imbalance. We showed the efficacy of our method using the problem of Arabic PAWs classification. In the future we would also use the data bias with classifier bias by exploring more features of the data with this approach.

REFERENCES

[1] K. Daifallah, N. Zarka, H. Jamous, *Recognition-Based Segmentation Algorithm for On-Line Arabic Handwriting*, In Proceedings of International Conference on Document Analysis and Recognition. ICDAR '09. pp. 886-890

[2] El-Sheik and El-Taweel, *Real-Time Arabic Handwritten Character Recognition*, In proceedings of Pattern Recognition, volume 23 (1990), No 12, pp. 1323-1332

[3] P. Natarajan, K. Subramanian, A. Bhardwaj and R. Prasad, *Stochastic Segment Modeling for Offline Handwriting Recognition*, In proceedings of International Conference on Document Analysis and Recognition. ICDAR '09. pp. 971-975

[4] T.G. Dietterich, *Ensemble Methods in Machine Learning*, In Proceedings of First International Workshop on Multiple Classifier Systems. Newyork, 2000

[5] L.M. Lorigo and V. Govindaraju, *Offline Arabic Handwriting Recognition: A Survey*, In Proceedings of Pattern Analysis and Machine Intelligence, 2006. pp. 712-724

[6] R. Khoussainov, A. He, N. Kushmerick, *Ensembles of biased classifiers*, In Proceedings of International Conference on Machine Learning, 2005. pp. 425-432

[7] T.M. Mitchel, *Machine Learning*, McGraw-Hill, 1997

[8] 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.

[9] J. Chen, H. Cao, R. Prasad, A. Bharadwaj and P. Natarajan, *Gabor features for offline Arabic handwriting recognition*, In Proceedings of International Workshop on Document Analysis Systems. DAS,2010.

[10] A. AbdulKader *Two-Tier Approach for Arabic Offline Handwriting Recognition*, In Proceedings of Conference on Arabic and Chinese Handwriting Recognition, 2006. pp. 70-81