2011 International Conference on Document Analysis and Recognition

A Semi-supervised SVM framework for Character Recognition

Amit Arora and Anoop M. Namboodiri
Center for Visual Information Technology, IIIT, Hyderabad, INDIA - 500 032.
{*amit.arora@research., anoop@*}*iiit.ac.in*

***Abstract*—In order to incorporate various writing styles or fonts in a character recognizer, it is critical that a large amount of labeled data is available, which is difficult to obtain. In this work, we present a semi-supervised SVM based framework that can incorporate the unlabeled data for improvement of recognition performance. Existing semi-supervised learning methods for SVMs work well only for two-class problems. We propose a method to extend this to large-class problems by incorporating a participation term into the optimization process. The proposed system uses a Decision Directed Acyclic Graphs (DDAG) of SVM classifiers, which have proven to be very effective for such recognition problems. We present experimental results on three different digits dataset with varying complexity, as well as additional multi-class datasets from the UCI repository for comparison with existing approaches. In addition we show that approximate annotations at the word or sentence level can be used for evaluation as well as active learning to further improve the recognition results.**

***Keywords*-Semi-Supervised SVM, Decision Directed Acyclic Graphs, Character Recognition**

I. Introduction

A large high quality labeled dataset that covers all sample variations within classes that are observed in the real world is critical for the development of an accurate recognition engine such as for handwritten or printed characters. However, such datasets are extremely difficult and costly to build, primarily due to the cost of accurate annotation of samples that cover all possible variations that need to be handled by the recognizer. The problem is more acute for South Asian languages, where the number of character classes is relatively large. Moreover, high quality recognizers are not available for most of these languages, commercially.

Support vector machines (SVMs) are commonly employed in many such recognition problems due to its generalization power. While the generalization power of SVMs allows us to generate recognizers with reasonable accuracies from fewer number of labeled samples, it will not be able to model all the writing/font styles or exactly map decision boundaries between classes unless labeled samples available show the regions of separation.

One of the common solutions that has emerged to handle the lack of annotated data in learning problems is Semi-supervised learning, which tries to make use of large quantities of unlabeled samples along with a small quantity of annotated samples to learn the classifier. The spatial coherence of the unlabeled samples are used to refine the decision boundaries within the recognizer.

Table I
Applying semi-supervised SVM training directly on 3 digit datasets. Cross-validation results with 10% labeled data.

Dataset	Only labeled data	Semi-supervised SVM
Pen Digits	91.8% (±1.2)	85.9% (±2.3)
Opt Digits	86.7% (±2.4)	82.5% (±2.0)
Online HW Digits	90.9% (±0.6)	85.4% (±1.3)

However, there is a significant problem that hampers the use of semi-supervised techniques for SVM. These methods are primarily designed for two-class problems, where the unlabeled samples are assumed to belong to one of the two classes under consideration. The assumption of two class problem is not such a hinderance as in most cases, as we often use a combination of multiple one-vs-one classifiers to create a multi-class classifier. In our experiments, Decision Directed Acyclic Graphs (or DDAGs) have been used for the combination. However, for problems such as character recognition, where the number of classes are relatively large, the assumption that each unlabeled sample belongs to one of the two classes under consideration is not valid. As we can notice from Table I, a direct application of semi-supervised SVM (S3VM) learning to such problems will result in a reduction in accuracy as compared to using just the labeled samples for training. In each case, 10% of the data was treated as labeled and the remaining as unlabeled for training, and the results presented are based on 10-fold cross validation, along with standard deviation across trials.

As noted above, the primary factor that affects the use of Semi-Supervised SVM training in such problems is the assumption about unlabeled data that is made during the learning process. The S3VM techniques assume that each unlabeled sample belongs to one of the two classes under consideration. As we see from the table, the accuracy of recognition often reduces due to the use of unlabeled samples from other classes.

In this work, we develop a framework for training two-class SVMs in datasets that contain multiple classes using Semisupervised SVM training. The primary approach is to attach a participation term p_i to each sample for each two-class SVM that we train. The participation term for unlabeled samples depends on the distribution labeled samples around it in the kernel space. The

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.223

influence of unlabeled samples are modulated by p_i and is included in the optimization process. In many text/handwriting datasets, it is often easier to obtain a coarse annotation of the dataset at a sentence or word level as opposed to character or symbol level. We also suggest a method to make use of such coarse annotation to improve the recognition performance, even further.

A. Previous Work

As cited in Chapelle *et al.* [1], and Zhu *et al.* [2], Semi-supervised methods for classification are based on a semi-supervised smoothness assumption, cluster assumption or a manifold assumption. The major semi-supervised learning algorithms can be categorised as: graph-based, boosting-based and density-based.

The graph-based methods predict class labels based on neighborhood, so that they are smooth on graph of unlabeled examples. These approaches use different parameters to define the smoothness of class labels. These algorithms are mostly restricted to solve two-class problems. Some of the algorithms using this method are Minimum Cuts [3], Harmonic Functions[4] and Manifold Regularization [5].

SemiBoost [6] is a boosting method based on cluster and manifold assumption, but solves two-class problems. MCSSB [7] presents an approach for boosting for multi-class classification. This method although is efficient for lesser number of classes, but the accuracy decreases for large class problems(≥ 10 classes).

The semi-supervised SVMs [8], [9], [10] are based on cluster assumption, and hence build hyperplanes at places of least density to separate two classes. A multi-class method based on semi-definite programming has been presented in [11], but it has a high computational cost attached to it.

II. Multi-class SVMs for Character Recognition

Multi-class SVMs are often built using a combination of multiple two-class SVMs. Popular approaches for this purpose includes building one-versus-rest classifiers for each class; using majority vote on one-versus-one classifiers between each pair; using a binary tree structure using half-versus-half classifiers [12]; and a Decision Directed Acyclic Graph (DDAG) structure using one-versus-one classifiers between each pair [13]. In our experience, the one-versus-one approaches are most accurate when used for large-class problems such as character recognition [14]. Unfortunately, those approaches are exactly the ones that violates the assumptions in Semi-supervised SVM training. We stick DDAG based combination of SVMs for our experiments, although the proposed approach is applicable to majority voting based classification as well.

We will now take a look at the Semi-supervised SVM training process and the proposed modification to deal with unlabeled samples from multiple classes in the training process.

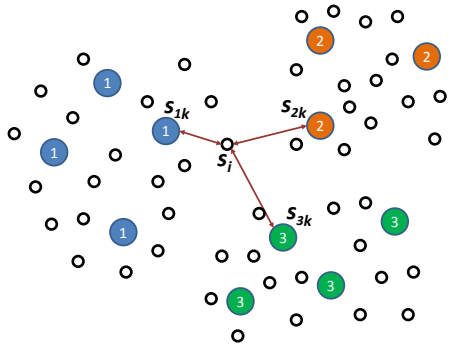

Figure 1. An unlabeled sample S_i and its nearest labeled samples S_{1i}, S_{2i}, and S_{3i}, from classes ω_1, ω_2, and ω_3 respectively.

A. Semi-supervised Multi-class SVM Training

Consider a dataset of c classes, $\mathcal{D} = \{D_1, D_2, \ldots, D_c\}$, where $D_i = \{L_i, U_i\}$ represents the subset of samples from class ω_i. L_i denotes the set of labeled samples and U_i denotes the set of unlabeled samples from class ω_i. Note that during the training process the unlabeled samples of all classes together form a single set, while we assume the labels for testing purposes only.

Consider a set of labeled and unlabeled samples belonging to three classes in Figure 1. The labeled samples are shown in large solid circles with label and the unlabeled ones are the smaller empty circles. Consider the unlabeled sample S_i in the above figure. The similarity of the sample S_i to samples in different classes such as S_{1k}, S_{2k}, and S_{3k} can be computed using a similarity measure, $Sim(S_i, S_{jk})$. We can extend this to define a similarity measure, $m_{i,j}$, between a sample S_i and class ω_j as:

$$m_{i,j} = \min_k Sim(S_i, S_{jk}). \tag{1}$$

We can define any measure of similarity, $Sim() \in [0, 1]$, based on quantities such as the dot product in the kernel space, inverse of Euclidean distance, etc., for this purpose.

The primary problem that we face in dealing with the unlabeled samples is that they can belong to any of the c classes, and hence we do not know whether to include a particular unlabeled sample while learning the two-class classifier ω_i vs, ω_j. To overcome We define the participation term p_k for each sample S_k. The value of $p_k(\omega_i, \omega_j)$ for the ω_i $vs.$ ω_j classifier is defined as:

$$p_k(\omega_i, \omega_j) = \begin{cases} 1 & \text{if } S_k \in L_i \text{ or } L_j \\ 0 & \text{if } S_k \in L_{m \neq i,j} \\ \dfrac{min(m_{k,i}, m_{k,j})}{1 + \min\limits_{l \neq i,j} m_{k,l}} & \text{otherwise} \end{cases} \tag{2}$$

Consider the problem of learning a two-class classifier: ω_i $vs.$ ω_j. The participation term tends to be high for samples which are close to ω_i or ω_j, while being far away from labeled samples in other classes. Such samples contribute significantly to the computation of margin in the semisupervised SVM learning framework. Conversely, samples that are farther away from the classes under consideration and closer to other classes will have a lower participation term, and will not significantly affect the margin computation.

We now look at the formulation of the Semisupervised SVM training by incorporating the participation term.

B. SVM Formulation and Analysis

To formulate the SVM learning, we modify the penalty associated with each sample by multiplying with the participation term. The cost of misclassification during the learning process is hence reduced for samples that have a lower participation term. The optimization function remains the same, and the participation terms appear only as constraints in the optimization.

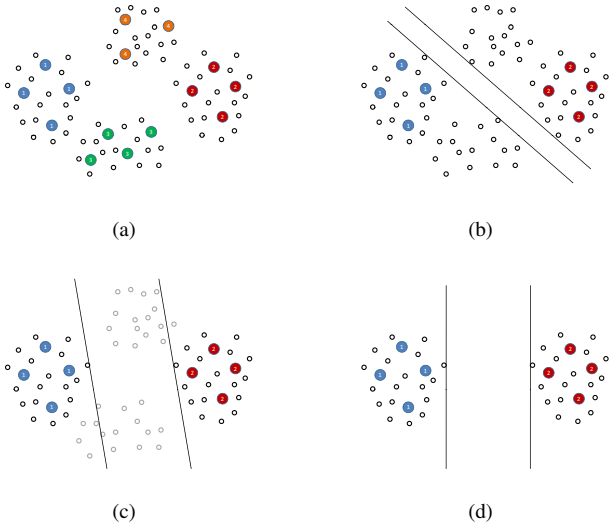

Figure 2. (a) A four-class problem, where we consider learning of the 1-vs.-2 classifier. The inclusion of all unlabeled samples in the learning of 1-vs.-2 results in S3VM producing the result in (b). The participation terms will reduce the significance of unlabeled samples from classes 3 and 4. The ideal case if we know the exact participation is shown in (d).

A cost sensitive optimization problem for SVM [15], where each sample i has a cost $p^{(i)}$ attached, as defined in equation (2), is formulated as

$$\begin{aligned} min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{l} p^{(i)}\xi_i \qquad (3)\\ \text{s.t. } y^{(i)}(w.x^{(i)}+b) \;&\geq\; 1-\xi_i \\ \xi_i \;&\geq\; 0 \end{aligned}$$

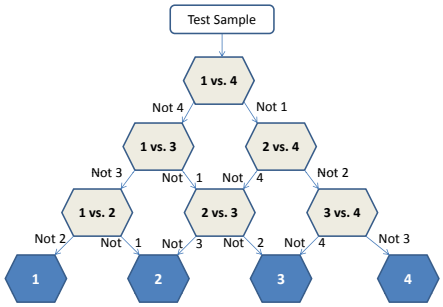

Figure 3. Classification using DDAG tree.

A semi-supervised formulation for SVM can be extended using participation cost from equation (3) in [9].

$$\begin{aligned} min_{w,b,\xi,\xi^*,y^*} \frac{1}{2}|w|^2 \;+\; C\sum_{i=1}^{l} p^{(i)}\xi_i + C^*\sum_{j=l+1}^{l+u} p^{(j)}\xi_j^* \\ \text{s.t. } y^{(i)}(w.x^{(i)}+b) \;&\geq\; 1-\xi_i \qquad (4)\\ y^{(j)*}(w.x^{(j)}+b) \;&\geq\; 1-\xi_j^* \\ \xi_i \;&\geq\; 0 \\ \xi_j^* \;&\geq\; 0 \end{aligned}$$

Here first l samples are labeled and there are u unlabeled samples. We need to find a labeling y^* for each of the unlabeled example, by using the cost-parameter C^* and slack variables ξ_j^*.

So, for training each binary classifier, a participation cost is assigned to each of the unlabeled example. The examples with very low participation cost are neglected as they normally belong to other classes in a multi-class scenario. The resultant SVMs can then be used in a DDAG structure for classification.

C. DDAG Architecture

For a multi-class problem, with n classes, 1-vs-1 classifiers are built for each pair resulting in $n*(n-1)/2$ such classifiers. A tree is built with these classifiers, which at each level decides that a sample doesn't belong to a particular class. Thus at the end of $n-1$ comparisons, class of a given sample is determined. A DDAG tree for 4-class system is shown in Figure 3. For such a formulation to work, the individual classifiers need to be independent of samples belonging to other classes, which we are trying to achieve with the SVM formulation described in the previous subsection.

Transductive SVM of svmlight [16] was modified to include participation term for unlabeled examples and DDAG classification structure was built over it.

To illustrate the effect of the participation terms, let us consider the four-class scenario described in Figure 2. The participation term reduces the effect of unlabeled samples

Table II
DESCRIPTION OF DATA SETS.

Data set	# Samples	# Attributes
Pen Digits	3498	16
Opt Digits	1797	64
Online HW Digits	3232	48

from other classes, leading to better separation between classes and generalization performance.

III. EXPERIMENTAL RESULTS AND ANALYSIS

We conducted several experiments to evaluate the effectiveness of the proposed solution. Three different datasets of offline and online handwritten digits were used for this purpose. All the datasets contain 10 classes (0-9), and their description is given in Table II. For cross validation trials, in each trial 10% of the data was separated out and treated as labeled, while the remaining was treated as unlabeled. Note that for all the tests, the complete dataset was used for testing, repeating the training process 20 times with different training sets. The average accuracies and standard deviations are reported in Table III. The first row indicates the results if all the data were used for training. Note that the accuracies in this case are based on resubstitution, and hence optimistic. The second row shows the results if we use only 10% of the data for training. This is the base for semi-supervised learning. The third row indicates the use the remaining 90% of the data as unlabeled, and carry out the traditional Semi-supervised SVM (S3VM) training. The results seem to degrade as we noted at the beginning of the paper.

The fourth row indicates the result of the proposed method (S3VM + Participation). We can clearly see the improvement in results from using just 10% of data for training. This demonstrates that the proposed method is able to effectively utilize the unlabeled data even for large multi-class problems. To test the limits of the approach, we check the maximum accuracy that is achievable using Semi-supervised SVM. This is done by assuming that we have perfect knowledge as to which all classifiers, each unlabeled sample should contribute. In other words, we assume that we have an ideal participation term that gives a 0 or 1 value depending on whether the sample is part of the pair of classes under consideration or not. The last row indicates the results we can achieve if such an ideal function is available. Note that our results using the proposed participation term is close to the one using the ideal participation function.

As mentioned in the previous work section, the work that most closely relates to the current work is by Valizadegan *et al.* [7]. We now compare their results with that of our algorithm using the same datasets that were reported in their paper (see Table IV). Note that the classifier used is different in both cases. We note that the proposed approach is able to achieve very good results in comparison with the MCSSB technique. We also note that the base accuracy of the DDAG-SVM classifier using 10% of data for training is sometimes higher than what is achievable using MCSSB. In most cases, our approach is able to improve on these results considerably.

Table III
SEMI-SUPERVISED SVM TRAINING WITH AND WITHOUT PARTICIPATION TERM. THE LAST ROW INDICATES THE RESULTS IF WE KNOW EXACTLY WHICH UNLABELED SAMPLE SHOULD PARTICIPATE IN EACH CLASSIFIER.

Training Dataset	Pen Digits	Op Digits	Online Digits
Completely Labeled	98.9% (± 0)	98.8% (± 0)	95.9% (± 0)
10% Labeled	91.8% (± 1.2)	86.7% (± 2.4)	90.9% (± 0.6)
S3VM	85.9% (± 2.3)	82.5% (± 2.0)	85.4% (± 1.3)
S3VM+ Participation	94.4% (± 2.0)	90.6% (± 2.5)	92.8% (± 0.5)
S3VM+ Ideal Participation	95.1% (± 1.7)	93.9% (± 2.0)	93.3% (± 0.6)

For the Balance and Car datasets, the proposed approach in fact shows a reduction in accuracy from using only 10% of data for training. This is primarily because of the presence of categorical data in the features, and hence the participation term does not give a good accuracy. In such cases, the MCSSB algorithm clearly performs the best. The yeast dataset contains a highly skewed distribution of samples within classes, and hence does improve the accuracy even with traditional S3VM training. The dermatology dataset contains a single nominal attribute along with other integral feature values, and our algorithm is able to deal with such data very well.

The third experiment was very specific to test the use of coarse annotation for improving the results. Even though this approach is applicable more to textual data than digits, we carry out a similar experiment by providing labels for a set of digits. The resulting classifier was able to locate a cluster of the digit 5 which was incorrectly classified, and the labeling of a single sample from this cluster improves the accuracy of the recognition of digit 5 by over 20%.

IV. CONCLUSION

In this work, we have presented an approach to use Semi-supervised SVM training for large-class recognition problems such as character recognition. The notion of participation of an unlabeled sample in a two-class classifier is introduced and quantified, and a method to integrate the participation term into the optimization process of semi-supervised SVM training is described. Experimental results on different character recognition datasets show the effectiveness of the approach in utilizing the unlabeled data. We also show that a coarse annotation of the data can be used along in an active learning fashion to improve the accuracy

Table IV
COMPARISON OF PROPOSED METHOD WITH MCSSB-MLP AND MCSSB-DT.

Dataset	10% labeled	S3VM	Proposed	MCSSB-MLP	MCSSB-DT
Pen Digits	91.8% (±1.2)	85.9% (±2.3)	94.4% (±2.0)	52.2% (±1.4)	57.7% (±1.2)
Opt Digits	86.7% (±2.4)	82.5% (±2.0)	90.6% (±2.5)	27.6% (±1.0)	33.9% (±1.3)
Iris	93.5% (±2.0)	75.5% (±7.9)	95.7% (±1.4)	84.1% (±2.3)	79.7% (±2.7)
Segmentation	79.1% (±4.4)	55.4% (±2.4)	80.4% (±4.6)	46.8% (±1.7)	48.5% (±2.3)
Dermatology	79.6% (±2.2)	88.5% (±4.6)	93.9% (±3.9)	65.6% (±2.1)	78.4% (±1.4)
Wine	92.1% (±3.6)	86.6% (±2.9)	96.9% (±1.4)	83.2% (±1.2)	83.2% (±1.2)
Yeast	42.6% (±2.7)	50.4% (±1.8)	54.4% (±1.6)	41.6% (±1.1)	47.8% (±0.6)
Balance	67.5% (±3.5)	73.9% (±7.2)	67.6% (±5.2)	86.6% (±0.6)	69.5% (±1.0)
Car	66.5% (±13.8)	77.2% (±5.3)	75.3% (±6.7)	78.0% (±0.4)	83.7% (±0.5)

of the classifier even further.

ACKNOWLEDGMENT

The authors would like to acknowledge the Technology Development for Indian Languages (TDIL) effort from the Ministry of Communication and Information Technology (MCIT) for funding part of the research in this work.

REFERENCES

[1] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, Sep. 2006.

[2] X. Zhu, "Semi-Supervised Learning Literature Survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep., 2005.

[3] A. Blum and S. Chawla, "Learning from Labeled and Unlabeled Data Using Graph Mincuts," in *Proc. International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 19–26.

[4] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. International Conference on Machine Learning*, 2003, pp. 912–919.

[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, Dec. 2006.

[6] P. Mallapragada, R. Jin, A. Jain, and Y. Liu, "Semiboost: Boosting for semi-supervised learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2000–2014, Nov. 2009.

[7] H. Valizadegan, R. Jin, and A. K. Jain, "Semi-supervised boosting for multi-class classification," in *Proc. European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ser. ECML PKDD '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 522–537.

[8] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 368–374.

[9] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 200–209.

[10] T. D. Bie and N. Cristianini, "Convex methods for transduction," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2003, pp. 73–80.

[11] L. Xu and D. Schuurmans, "Unsupervised and semi-supervised multi-class support vector machines," in *Proceedings of the 20th national conference on Artificial intelligence - Volume 2*. AAAI Press, 2005, pp. 904–910.

[12] H. Lei and V. Venu Govindaraju, "Half-against-half multi-class support vector machines," in *Proc. Multiple Classifier Systems*, 2005, pp. 156–164.

[13] T. K. Chalasani, A. M. Namboodiri, and C. Jawahar, "Support vector machine based hierachical classifieds for large class problems," in *Proc. Six International Conference on Advances in Pattern Recognition*, Kolkatta, 2007.

[14] N. Neeba, A. M. Namboodiri, C. Jawahar, and P. Narayanan, *Guide to OCR for Indic Scripts Document Recognition and Retrieval*. Springer, 2009, ch. Recognition of Malayalam Documents, pp. 125–146.

[15] P. Geibel, U. Brefeld, and F. Wysotzki, "Perceptron and svm learning with generalized cost models," *Intell. Data Anal.*, vol. 8, pp. 439–455, October 2004.

[16] T. Joachims, *Making large-scale support vector machine learning practical*. Cambridge, MA, USA: MIT Press, 1999, pp. 169–184.