Co-Training with Adaptive Bayesian Classifier Combination

Yusuf Yaslan [1]and Zehra Cataltepe [2]

Istanbul Technical University Computer Engineering Department
34469 Maslak, Istanbul/Turkey
[1]yyaslan@itu.edu.tr [2]cataltepe@itu.edu.tr

***Abstract*— In a classification problem, when there are multiple feature views and unlabeled examples, co-training can be used to train two separate classifiers, label the unlabeled data points iteratively and then combine the resulting classifiers. Especially when the number of labeled examples is small due to expense or difficulty of obtaining labels, co-training can improve classifier performance. For binary classification problems, mostly, the product rule has been used to combine classifier outputs. In this paper, we propose an adaptive Bayesian classifier combination method which selects either the Bayesian or the product combination method based on the belief values. We compare our adaptive Bayesian method with Bayesian, product and maximum classifier combination methods for the multi-class pollen image classification problem. Two different feature sets, Haralick's texture features and features obtained using local linear transforms are used for co-training. Experimental results show that adaptive Bayesian combination with co-training performs better than the other three methods.**

I. Introduction

In many pattern recognition applications, in addition to labeled training data, unlabeled data is also available. The unlabeled data becomes available where obtaining the inputs for data points is cheap, however labeling them is time, money and effort consuming. For example, in speech recognition, recording huge amount of audio doesn't cost a lot. However, labeling it requires someone to listen and type. Similarly, billions of web pages can be obtained from web servers. However, classifying these web pages into classes is a time consuming and difficult task. Similar situations are valid for remote sensing, face recognition, medical imaging and intrusion detection in computer networks [8]. Semi-supervised learning methods [1] are used in order to make use of unlabeled data. In some applications, data samples obtained from various sources may be represented in different multiple ways (or views), for example, web pages can be represented by using both textual information and hyperlink structure information between them [5]. Generally, when there is more than one feature view, they are concatenated to form the whole feature space. However this may sometimes be problematic, i.e. the concatenated features may lack physical meaning [7]. These different views can also be used for training more than one classifiers. The co-training algorithm was proposed to reduce the misclassification rate by reducing the disagreement between classifiers generated for different views [5].

One of the important steps in co-training is the classifier combination, where the predicted labels on each feature set are combined to produce a final labeling. Product rule is the combination method that has been used most of the time. In this study we explore the Bayesian [9] and maximum [3] classifier combination techniques for co-training and in addition propose an adaptive Bayesian combination rule.

Experimental results are obtained on Pollen image dataset. Pollen analysis is important in the study of allergic reactions, search for hydrocarbons in medicine, derivation of geographical origin of products [11], paleo-ecology and paleo-climatic reconstruction [12]. The tasks of classification of pollen grains are laborious and require highly skilled people.

Previously pollen classification was studied by [11], [14], [15] for three types of pollen of the Urticaceae family. In [14] area, perimeter, compactness, centroid, mean distance to centroid, maximum distance to centroid, minimum distance to centroid and diameter features based on shape are used by minimum distance classifier. Brightness and shape based descriptors are also used as pollen features [11]. In a detailed study [16], Haralick's coefficients, gray level run length statistics, local linear transformations, neighboring gray level dependence statistics, first-order statistics, energy and entropy features computed for three levels of decomposed wavelet packets are evaluated by Support Vector Machine, K-Nearest Neighbour and Multi-Layer Perceptron classifiers.

In this paper, unlike the previous studies, we use a co-training algorithm for pollen image classification. Two different feature splits from co-occurrence matrix and local linear transform are obtained. Logistic Linear Classifier is used as the classifier of co-training algorithm. In co-training, usually product combination is used for combining the output probabilities. We propose an adaptive Bayesian classifier combination, which eliminates the zero belief values that may be obtained due to small amount of labeled data. We compare the classification accuracy results of Adaptive Bayesian with Bayesian, product and maximum product rules. We show that co-training helps with classification of pollen patterns and adaptive Bayesian combination gives the best results.

II. Co-Training Algorithm

When there are more than one feature splits for data, they can each be used to train a classifier. Co-training algorithm is

an iterative algorithm, proposed to train classifiers on different feature splits and compensate each others' classification error by adding the most surely classified data samples from unlabeled data. Under certain assumptions, by starting with a weak classifier co-training algorithm can learn from the unlabeled data. The first assumption is that the target function over each feature set predicts the same label (compatibility). The second assumption is, given the class of the instance, the feature sets are conditionally independent [5]. It is, however, difficult for real datasets to satisfy compatibility and conditional independence. In the general co-training algorithm the feature sets are referred to as views and it is assumed that two different views such as F_1 and F_2 are available. The overall feature set F is the concatenation of the different views: F = F_1 U F_2. The general co-training algorithm starts with a set of labeled data L and unlabeled data U. Than creates a pool U' by choosing u examples at random from U. The algorithm iterates a specified number of items and does the following: By using L it trains classifiers C_1 and C_2 that considers only the F_1 and F_2 portion of F respectively. C_1 and C_2 label examples from U' and select the most surely classified single example for each class. Each classifier adds self-labeled examples to L. Than the algorithm randomly chooses examples from U to replenish U'. The block diagram of the co-training system is given in fig.1.

Two classifiers, C_1 and C_2, predict class labels for data samples. At each iteration, we select the samples from U', if a classifier is sure about that sample above a threshold. This process is continued until the number of data samples in U' are less than a threshold. Then the predictions are combined. Most of the previous research combined the predictions by multiplying their class probability scores together and then renormalizing them. In this work, we use new classifier combination methods for co-training algorithm.

PrTools [18] implementation of Logistic Linear Classifier is used [4] as the base classifier for co-training algorithm. Previously it was used for medical image analysis [2] in a co-training approach. The combination scheme was naive Bayes and the proposed method in [2] also considered hand labeling the data samples. However we don't get require an oracle and use Bayesian classifier combination. In the experiments one against all classification scheme is used.

A. Classifier Combination for Co-training

Let m be the number of classes, w_i be the ith class label, R be the number of classifiers, x_i be the measurement vector used by the ith classifier and Z be the pattern to be classified. Given measurements x_i, $i = 1, 2, ..., R$, the pattern Z is assigned to class w_j, provided the posterior probability of that interpretation $P(w_j|x_1, ...x_R)$ is maximum [3].

In order to reach a decision, probabilities of various hypotheses should be computed by considering all features. Using the Bayes theorem we can write the $P(w_k|x_1, ...x_R)$ as:

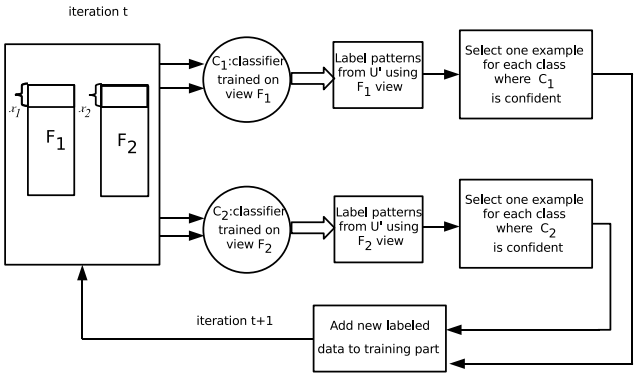

Fig. 1. Block diagram of co-training algorithm.

$$P(w_k|x_1, ...x_R) = \frac{P(x_1, ...x_R|w_k)P(w_k)}{P(x_1, ...x_R)} \quad (1)$$

$P(x_1, ...x_R)$ (the unconditional feature joint probability density) can be expressed in terms of conditional feature distributions:

$$P(x_1, ...x_R) = \sum_{j=1}^{m} P(x_1, ...x_R|w_j)P(w_j) \quad (2)$$

We consider three different rules for classifier combination in co-training.

1) Product Rule: $P(x_1, ..., x_R|w_k)$ represents the joint probability distribution of the features extracted by the classifiers. Assume that the representations are conditionally statistically independent. Then we can rewrite $P(x_1, ..., x_R|w_k)$ as:

$$P(x_1, ..., x_R|w_k) = \prod_{i=1}^{R} P(x_i|w_k) \quad (3)$$

We can rewrite the posterior probability as:

$$P(w_k|x_1, ..., x_R) = \frac{P(w_k)\prod_{i=1}^{R} P(x_i|w_k)}{\sum_j^m P(w_j)\prod_{i=1}^{R} P(x_i|w_j)} \quad (4)$$

Using the Bayes rule results in:

$$P(w_k|x_1, ..., x_R) = \frac{\prod_{i=1}^{R} P(w_k|x_i)/P(w_k)^{R-1}}{\sum_{k'}\{\prod_{j'} P(w_{k'}|x_{j'})/P(w_{k'})^{R-1}\}} \quad (5)$$

Product rule assigns pattern Z to the class J which maximizes the right handside of Eq.5.

If a priori class probabilities are equal ($P(w_j)$ =1 / (number of classes)) this formula reduces to product combination.

2) Max Rule: Under the assumption of equal priors and conditional independence given class labels the max rule assigns pattern Z to class w_j if [3]

$$J = arg \max_{k=1}^{m} \max_{i=1}^{R} P(w_k|x_i) \quad (6)$$

3) Bayesian Rule: Bayesian combination rule takes into consideration each classifier's performance. The performance of a classifier is indicated by it's confusion matrix C, where C_{ij} denotes the number of patterns with actual class i, classified as class j by the classifier. Total number of patterns that are classified as class j can be obtained by $\sum_{i=1}^{m} C_{ij}$. The conditional probability that a pattern x actually belongs to class i given that classifier r assigns it to class j can be estimated as [9]:

$$P(x \ \epsilon \ C_i | e_k(x) = j) = \frac{C_{ij}^{(r)}}{\sum_{i=1}^{m} C_{ij}^{(r)}} \tag{7}$$

where $C_{ij}^{(r)}$, $1 \leq r \leq R$, represents the ith row and jth column of rth classifier's confusion matrix. $e_k(x)$ is classifier k's decision. Eq.(7) represents the degree of accuracy when classifier r assigns class i to a pattern.

Let $e_r(x) = j_r$ for $1 \leq r \leq R$ be the classification results of any pattern x obtained by R classifiers, then a belief value that x belongs to class i can be defined as:

$$bel(i) = P(x \ \epsilon \ C_i | e_1(x) = j_1, ..., e_R(x) = j_R) \tag{8}$$

Assuming independence of classifiers and applying Bayes formula belief value can be approximated as [9]:

$$bel(i) = \frac{\prod_{r=1}^{R} P(x \ \epsilon \ C_i | e_r(x) = j_r)}{\sum_{i=1} m \prod_{r=1}^{R} P(x \epsilon \ C_i | e_r(x) = j_r)} \tag{9}$$

for $1 \leq i \leq m$. Input pattern x is assigned to class j if $bel(j) > bel(i)$ for all $i \neq j$.

In our experiments we computed confusion matrices for F_1, F_2 and F feature sets. Considering the fact that, small amount of data samples in the training set may lead to zero belief values. Due to sparsity of labeled data when co-training, we used the training set to compute the confusion matrices. We propose to use an adaptive Bayesian combination scheme for co-training. If the maximum belief value for a data sample is less than a threshold, instead of Bayesian rule, we use product combination for that data sample. Experimental results show that this adaptive combination of co-training improves the classification performance.

III. Test Results on Pollen Image Dataset

A. Data Set

In this work seven types of pollen images from Bangor/Aberystwyth Pollen Image Database [13] is used. The types of pollen cells belongs to the following classes: Plantago lanceolata, Quercus robor, Alnus glutenosa, Polypodium vulgare, Rumex acetosella, Conopodium majus and Dactylis glomerata. The dataset consists of relatively low spatial resolution images (typically 80-100 pixels in each dimension). The Polypodium vulgare type consist of 196 images and the others have more. Fig.2 shows different types of pollen images.

Previously, this dataset is used in [10] and the best classification performance of 83% obtained by using all features together and RBF networks as the classifier.

Fig. 2. Different types of pollen images from three different classes: lantago lanceolata, Quercus robor and Alnus glutenosa respectively.

B. Feature Sets

In this paper, in order to conserve the physical meanings of features, we use semi-supervised approach on two kinds of features. The first type of feature set is Haralick's texture features [6] that uses co-occurrence Matrix (CM) obtained form each image and the second one uses Local Linear Transforms (LLT) [17]. Haralick's texture features [6] uses the co-occurrence Matrix obtained form each image. The co-occurrence matrix is calculated using the relative distance among the pixels and their relative orientation. It captures a significant amount of textural information. Based on these matrices, Contrast, Inverse Difference Moment, Angular Second Moment and Entropy features are obtained for different orientations [6]. A total of 20 Haralick's features are extracted for each image.

Local linear transforms features (LLT) [17] are statistical measurements of the outputs of filter banks applied to images. Filters are designed to extract a particular feature from the texture of the image. A total of 9 features are obtained from local linear transforms. For details of these features please see [17].

In the results, Haralick's feature set is referred to as the feature set 1 and local linear transform based feature set is referred to as the feature set 2.

C. Experiment Details

In order to balance the dataset, for each class 196 images are used. Initially, the dataset is splitted into training and testing part with equal amount of data. 15% and 25% of the training data is used as labeled training data and the rest is used as the unlabeled training data. Co-training is used for 30 iterations (i.e. 30 unlabeled data points are labeled for each class). During the iterations the unlabeled data is divided into 2 equal random parts and one part is classified by logistic linear classifiers. At each iteration, the data points which are classified with a probability above a certain threshold are added into the labeled dataset. Experimental results are obtained for random 10 runs and the mean values of these runs are reported.

D. Results

In fig. 3 classification performance with co-training is given. Initially, 15 labeled samples are used for each class in the training phase. As shown in the figure, the Bayesian Combination method doesn't perform well because of the zero belief values. However adaptive Bayesian combination performs best. On the other hand classification performance

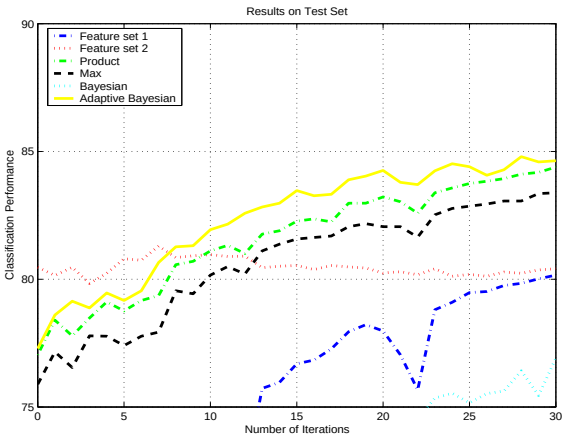

Fig. 3. Classification performance of linear logistic classifier on pollen image dataset. Initial labeled examples are 15 for each class.

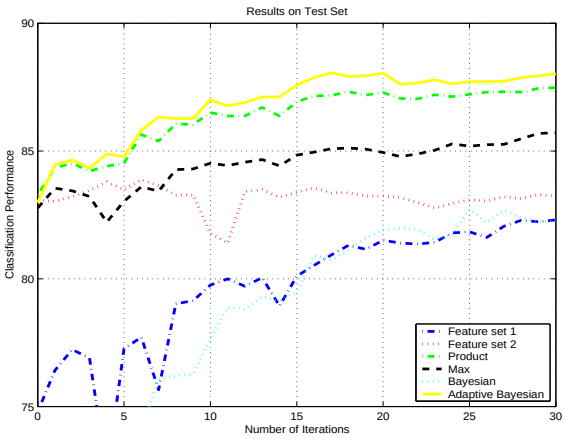

Fig. 4. Classification performance of linear logistic classifier on pollen image dataset. Initial labeled examples are 25 for each class.

for combination rules are less than performance of feature set 2 for initial iterations. At the end of the iterations, though, the product and maximum combination rules increase their performances, Adaptive Bayesian combination is still better than the other methods.

When the initial labeled samples for each class in the training phase is increased to 25. Increasing the number of labeled examples also increases the general performance. Bayesian Combination method improves it's performance but it is not better than any of the single feature sets. Adaptive Bayesian combination performs best.

In order to understand contribution of co-training we evaluated classifiers trained on 105, 175 and 686 (on feature portion F) training examples. We found out that supervised classification with 105, 175 and 686 training samples give 75.45% 79.94% and 90.58% respectively. Note that the results for 25 initial labeled examples for each class is around 88%. This accuracy is slightly less than the accuracy when we use all available data for training.

IV. Conclusion

In experiments performed on pollen image dataset, we have shown that when we apply co-training the general classification performance improves. Unlike the previous applications which combined the classifier outputs by multiplying class probabilites of each classifier, we apply adaptive Bayesian classifier combination scheme. Experimental results show that adaptive Bayesian combination with co-training gives best results.

References

[1] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-Supervised Learning, MIT Press,2006.
[2] W. He, X. Huang, D. Metaxas and X. Ying, "Efficient Learning by Combining Confidence-Rated Classifiers to Incorporate Unlabeled Medical Data", LNCS, Medical Image Computing and Computer-Assisted Intervention (MICCAI 2005) ,pp 745-752, 2005.
[3] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, "On Combining Classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.20, no.3, March 1998.
[4] A. Webb, Statistical Pattern Recognition, John Wiley & Sons, New York, 2002.
[5] A. Blum, T. Mitchell, "Combining Labeled and Unlabeled Data with Co-training" Proc. of the 11th Annual Conference on Computational Learning Theory (COLT '98) (pp. 92-100),1998.
[6] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, vol.3, pp 610-621, 1973.
[7] Y. Xu, C. Zhang and J. Yang, "Semi-Supervised Classification of Musical Genre Using Multi-View Features", International Computer Music Conference (ICMC 2005),5-9 September 2005.
[8] F. Roli, "Semi-supervised Multiple Classifier Systems: Background and Research Directions" Proc. of 6th Int. Workshop on Multiple Classifier Systems, 2005.
[9] C. Y. Suen and L. Lam "Multiple Classifier Combination Methodologies for Different Output Levels" Proc. of the 1st International Workshop on Multiple Classifier Systems, pp.52-66, 2000.
[10] F. Kesgin and Y. Yaslan, "Pollen Classification Using RBF Networks" IASTED International Conference on Computational Intelligence,San Francisco, USA, 2006.
[11] M.R-Damian, E. Cernadas, A. Formella, M.P.D. Sa-Otero, "Pollen classification using brightness-based and shape-based descriptors", Proc. 17th. International conference on Pattern Recognition (ICPR2004), Cambridge, England pp. 212-215, 2004.
[12] P. Li and J.R. Flenley, "Pollen Texture Identification Using Neural Networks", Grana, 38(1) pp. 59-64, 1999.
[13] I. France, A.W.G. Duller,G.A.T. Duller, and H.F. Lamb, "A new approach to automated pollen analysis", Quaternary Science Reviews 19 (6), pp. 537-546, 2000.
[14] M.R. Damian , E. Cernadas, A. Formella, and M.P.D. Sa-Otero, "Pollen Classification of Three Types of Plants of the Family Urticaceae", Proc. of the 12th Portuguese Conference on Pattern Recognition, Aveiro, Portugal, 2002.
[15] M.R. Damian, E. Cernadas, A. Formella, and A. Gonzalez, "Automatic Identification and Classification of Pollen of the Urticaceae Family", Proc. of Acivs 2003 (Advanced Concepts for Intelligent Vision Systems), Ghent, Belgium, September 2-5, 2003.
[16] M.F. Delgado, P. Carrion, E. Cernadas, and J.F. Galvez, "Improved Classification of Pollen Texture Images Using SVM and MLP", Proc of Visualization, Imaging, and Image Processing -(VIIP 2003) Benalmadena, Spain, pp. 232-237, 2003.
[17] A. Drimbarean and P.F. Whelan, "Colour Texture Analysis:A Comparative Study", Proc. Irish Machine Vision and Image Processing Conference 2000, The Queens University Ireland, pp. 125-132, 2000.
[18] R.P.W. Duin, "PRTOOLS A Matlab Toolbox for Pattern Recognition", 2004.