Semi-supervised Robust Alternating AdaBoost*

Héctor Allende-Cid[1], Jorge Mendoza[2], Héctor Allende[1,2],
and Enrique Canessa[2]

[1] Universidad Técnica Federico Santa María,
Dept. de Informática, Valparaíso-Chile
vector@inf.utfsm.cl
[2] Universidad Adolfo Ibáñez,
Facultad de Ingenieria y Ciencias, Viña del Mar-Chile
jorge.mendoza2003@alumnos.uai.cl, hallende@uai.cl, ecanessa@uai.cl

Abstract. Semi-Supervised Learning is one of the most popular and emerging issues in Machine Learning. Since it is very costly to label large amounts of data, it is useful to use data sets without labels. For doing that, normally we uses Semi-Supervised Learning to improve the performance or efficiency of the classification algorithms.

This paper intends to use the techniques of Semi-Supervised Learning to boost the performance of the Robust Alternating AdaBoost algorithm.

We introduce the algorithm RADA+ and compare it with RADA, reporting the performance results using synthetic and real data sets, the latter obtained from a benchmark site.

Keywords: Semi-Supervised Learning, Expectation Maximization, Machine ensembles, Robust Alternating AdaBoost.

1 Introduction

In supervised learning, classification tasks require training data with a class label. However, there are many real problems where the existence of labeled data is scarce and unlabeled data is abundant, either due to its cost or because it is difficult to obtain it (i.e. classification of text and web pages, processing medical imaging, diagnosis of industrial processes, speech recognition, protein structures, etc.). For this reason, it is necessary to build classifiers that work with a small amount of labeled data and a large amount of unlabeled data so they can learn from both. The main idea behind the algorithm RADA [1] (acronym for Robust Alternating AdaBoost) is to alternate the use of classical and inverse AdaBoost in order to lessen the impact of data outliers in the final classification.

In this paper we propose a generalization of the algorithm RADA for use in Semi-Supervised learning problems. Basically, the aim is to make use of the robust properties of the algorithm and extend it to take advantage of unlabeled

[*] This work was supported in part by Research Grant Fondecyt 1070220, and DGIP-UTFSM Grant.

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 579–586, 2009.

data to train the weak classifiers. In Section 2 we briefly introduce the algorithm RADA. Section 3 presents the analysis of the generalization of RADA to Semi-Supervised classification. In Section 4 we present the proposed algorithm RADA+. Experimental results are presented with both synthetic data and real data in Section 5. The last section is devoted to concluding remarks.

2 Robust Alternating AdaBoost

RADA is an acronym for Robust Alternating AdaBoost. As its name suggests, this algorithm combines the power of classical and inverse AdaBoost to reduce the impact of data outliers in training samples. The RADA algorithm bounds the influence of the outliers to the empirical distribution, it detects and diminishes the empirical probability of "bad" examples, and it performs a more accurate classification under contaminated data.

RADA computes the relative weight of each instance in the training set using a different equation. Originally AdaBoost obtains the relative weight as follows:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \tag{1}$$

RADA uses a robustified equation of α_t for smoothing the impact of an outlier data:

$$\alpha_t = \begin{cases} \frac{1}{2} \sqrt[r]{\ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)} + \alpha_\gamma & \epsilon_t < \gamma \\ \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right) & \epsilon_t \geq \gamma \end{cases} \tag{2}$$

where $\alpha_\gamma = \frac{1}{2} \ln \left(\frac{1-\gamma}{\gamma} \right) - \frac{1}{2} \sqrt[r]{\ln \left(\frac{1-\gamma}{\gamma} \right)}$ is a constant needed so that equation (2) is continuous.

Applying the previous equation will prevent the empiric distribution from growing considerably in one step for any sample. However, the empirical distribution is updated at each stage, and after a few iterations the probability weight of the samples that were misclassified repeatedly, will have bigger values compared to other samples. To solve this problem, Allende-Cid et al. [1] introduces two new variables to the algorithm: an inverting variable $\beta(i)$ and an age variable $age(i)$ for each example $i = 1 \ldots n$.

When the variable $\beta(i)$ value is 1, the algorithm behaves as the classical AdaBoost, i.e., the empirical distribution increases when a sample is misclassified and decreases the value otherwise. If the value of $\beta(i) = -1$, then the algorithm behaves like the Inverse AdaBoost, i.e., decreases the empirical distribution when a sample is misclassified and increases it otherwise. The variable $age(i)$ counts the number of times a sample i has been misclassified, if the number exceeds a threshold τ then the value of $\beta(i)$ is changed to -1 (originally the value of $\beta(i)$ for all samples is 1).

3 RADA Semi-supervised Generalization

The algorithm RADA compares the actual values of the class versus the value estimated by the ensemble to update the sampling distribution. Thus those instances that have been difficult to classify, i.e., the classification of the weak classifier differs from the actual class, will have a bigger probability to be selected in the training set on the next iteration. One of the main problems that arise from the method presented is that an unlabeled data has no "real" class to compare it with.

RADA algorithm defines the margin of an instance obtained in the i-th iteration by the equation

$$\alpha_T \beta(i) y_i h_T(x_i) \tag{3}$$

The problem, therefore, lies in defining the margin for an unlabeled data. To solve this problem, we take advantage of the cluster and manifold assumptions [5]. This seeks to improve the margin of classification (equivalent to minimize the error of the ensemble) through the selection of unlabeled data with a higher confidence rating, and assigns the class predicted by the current classifier.

To allow the same margin to be used for both labeled and unlabeled data, we use the same definition of *pseudo-class* as [4]. A subset of labeled data in addition to a group of unlabeled data and their pseudo-class, are used for training the weak classifier in the next iteration. This same strategy is used by algorithms such as ASSEMBLER [4], Self-Training [10] and Semisupervised MarginBoost [6].

First we find a mechanism to define the margin of an unlabeled data. For that we use the same function used in RADA, defining the base classifier as $h_T(x) : \mathbb{R} \rightarrow \{-1, 1\}$ where h_T is the T-th classifier in the ensemble. The set of training labeled data, L, is n-dimensional type of $x_1, x_2, \ldots x_n$ with their respective classes $\{-1, 1\}$. The classifier of the ensemble $H_T(x)$ is a linear combination of classifiers in step T

$$H_T(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{4}$$

where α_t is the equivalent weight in the algorithm RADA.

Now when adding a unlabeled data set, U, a margin associated with these data must be defined (as in the case of labeled examples). However, there is no knowledge of the true value of the class of the data, so following [6,3,8,11] we define the margin for an unlabeled data x_i as

$$|h_T(x_i)| \tag{5}$$

since the above expression is an absolute value, one can apply a mathematical simplification to represent this term

$$y_i h_T(x_i) \tag{6}$$

This allows to generalize the concept for both labeled and unlabeled data. If x_i is a labeled data, then y_i is the known class, on the contrary, if x_i is a unlabeled data, then y_i represents the pseudo-class (as defined above).

Once the problems of the margin were solved, we used the framework Expectation-Maximization (EM) [9]. EM is a popular iterative algorithm for maximum likelihood estimation in problems with unlabeled data. It consists two steps: the Expectation step and the Maximization step. The first step consists in classifying the unlabeled data based on the current hypothesis. Then the Maximization step re-estimates the parameters based on all the data (labeled and unlabeled with a pseudo-class). This leads to the next iteration of the algorithm, and so on. It has been proved that EM converge to a local minimum when the model parameters stabilize [9].

4 Semi-supervised Robust Alternating AdaBoost

In this section we present the proposed algorithm Semi-Supervised Robust Alternating AdaBoost (RADA+). The main idea of this proposal is to add unlabeled data, after a certain number of training epochs j, to the training data set in order to enhance the overall performance of the algorithm.

The developed framework is the following: At first we will take the labeled data and will train a non-ensemble based supervised classifier with it (in this particular case we took the SVM algorithm and trained it with the labeled data). After a number of j training epochs we compare the result of the H_j classifier with the SVM algorithm. If the result obtained from both of the classifiers is the same, we add these data examples to the training data set, hopefully to enrich the training phase of the algorithm. From the $j+1$ iteration on we use an EM approach to update the classification of the unlabeled data with the strong classifier H_{j+1}. The fundamentals behind our approach is to prove the impact of the strong classifier H_j on the weak classifiers. For this reason we propose to add the unlabeled data to the training data set in an iteration where the strong classifier is robust enough so as not to affect the final classification.

The parameters are defined in the following way:

$$D_j(x_U) = \max_{x \in \mathcal{U}} D_j(x), \quad age(i) = 0, \quad \beta(i) = 1$$

$D_j(x_U)$ is the initial weight of the unlabeled data examples when they are added to the training data set. We chose the maximum because we think that it is important that these examples are chosen when the resampling is made.

Algorithm 1 shows our proposed Semi-Supervised Alternating AdaBoost algorithm.

5 Experimental Results

In this section we empirically show the performance of our Semi-Supervised RADA (**RADA+**) model proposal compared to the **RADA** algorithm, for both

Algorithm 1. RADA+ Algorithm

```
Training Data Set $\mathcal{L} = \{(x_1, y_1), \ldots, (x_{n_l}, y_{n_l})\}$ with $n_l$ elements, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1, +1\}$
Unlabeled Data Set $\mathcal{U} = \{x_{n_l+1}, \ldots, x_{n_l+n_u}\}$ with $n_u$ elements, where $x_i \in \mathcal{X}$
Choose: $\tau$ age threshold, $\gamma$ limit threshold, the robust parameter $r$ and $t = 0$.
Train the non-ensemble based classifier with the training data set. Then perform a classification of the unlabeled data $\mathcal{U}$ with the classifier and assign them the corresponding pseudo-class $y$.
$D_1(i) = \frac{1}{n_l}$, $\beta(i) = 1$ and assign the $age(i) = 0$ variable to each sample $(x_i, y_i)$, $i = 1, \ldots, n_{n_l}$
repeat
   Increment $t$ by one.
   Select a bootstrap sample $Z_t$ from $\mathcal{Z}$ with distribution $D_t$.
   Construct $h_t : \mathcal{X} \to \{-1, +1\}$ classifier using the bootstrapped sample $Z_t$ as the training set.
   Calculate the ensemble error in step $t$:

      $$\varepsilon_t = \Pr_{i \sim D_t} [h_t(x_i) \neq y_i] = \sum_{i: h_t(x_i) \neq y_i}^{n} D_t(i)$$

   Calculate $\alpha_t$ as in (2).
   if $t = j$ then
      Classify the unlabeled data $\mathcal{U}$ with $H_{t-1}$.
      if $H_{t-1}(x_u) = y(x_u)$ then
         Add $x_u$ to the training set $\mathcal{Z}$
         Set the distribution $D_j(x_u) = \max_{x \in \mathcal{Z}} D_j(x)$, $age(x_u) = 0$ and $\beta(x_u) = 1$
      end if
   end if
   if $t \geq j + 1$ then
      Classify the unlabeled data that entered in the iteration $j$ with $H_{t-1}$
      Update the pseudo-class
   end if
   Update distribution

      $$D_t : D_{t+1} = \frac{D_t(i)}{W_t} \times e^{(-\alpha_t \beta(i) y_i h_t(x_i))}$$

   where $W_t = \sum_i D_t(i)$
   Final hypothesis $H_t$ in iteration $t$ is given by:

      $$H_t = sign\left(\sum_{k=1}^{t} \alpha_k h_k(x)\right)$$

   Classify $\mathcal{Z} = (x_1, y_1), \ldots, (x_n, y_n)$ with $H_t$
   if Sample $(x_i, y_i)$ was correctly classified by $H_t$ (meaning that $H_t y_i > 0$) then
      $age(i) = 0$ y $\beta(i) = 1$
   else
      Increment $age(i)$ by one
      If $age(i) > \tau$ then $\beta(i) = -1$ and $age(i) = 0$
   end if
until Stopping criterion is met
Output: hypothesis $H_t(x)$
```

Synthetic and Real data sets; the latter was obtained from the UCI Machine Learning Repository [2].

The data of both synthetic and real data sets were separated in labeled, unlabeled and test sets. The results reported for each model correspond to the mean value of the computed metrics, over 20 runs, using the same data sets.

For the synthetic data we used the following proportion labeled/unabeled data: 1%, 5% and 10%. For the real data sets, we used the following proportion labeled/unlabeled data: 5%, 10% and 20%. The difference lies on the data sets sizes, for the synthetic sets the total amount of data analyzed was 15000 instances, instead for the real data sets the amount of total data was approximately 5000 instances. The classifier used in the algorithms is the Bayesian Classifier (QDA) (see [7]). The non-ensemble based clasifier used to determine whether to add unlabeled data to the test set or not, was a Soft-margin Support Vector Machine with a Sequential Minimal Optimization method to find the separating hyperplane.

For the synthetic experiment we created a synthetic data set $\{(x_i, y_i)\}_{i=1}^{n}$, as an independent sample obtained from a mixture of gaussian distributions labeled with the class $\{-1, 1\}$. For more information on the details of the synthetic data sets, please refer to [1].

Table 1 shows the summary results of the performance evaluation on the synthetic data of the RADA and RADA+ algorithms, with 1%, 5% and 10% labeled data. As we can observe in the *Test Error* column, RADA and RADA+ have very similar behavior specially for the presence of a low percentage of labeled data. However, this radically changes when the amount of labeled data increases, i.e. 10%. Nevertheless RADA+ obtained good results in the training set, mainly because of the EM framework.

Table 1. Summary results of the performance evaluation of the RADA and RADA+ algorithms with 5% and 10% outliers

Labeled	Algorithm	Outliers	T	Train Error	Train Min.	Test Error	Test Min.
1%	RADA	5%	33.83	23.87 ± 10.44	15.19	**25.72 ± 9.48**	17.45
	RADA+	5%	15.8	**7.56 ± 11.52**	0.42	25.88 ± 4.52	20.14
	RADA	10%	32.1	26.71 ± 7.79	16.36	26.21 ± 7.12	16.81
	RADA+	10%	13.7	**8.04 ± 12.28**	0.50	**25.69 ± 4.05**	18.63
5%	RADA	5%	25.7	25.58 ± 1.55	24.16	25.56 ± 1.47	24.11
	RADA+	5%	20.2	**9.06 ± 10.49**	2.36	**24.52 ± 1.72**	22.25
	RADA	10%	15.1	23.36 ± 0.84	22.65	**23.87 ± 0.83**	23.14
	RADA+	10%	14.6	**8.60 ± 9.81**	2.28	25.07 ± 1.83	22.31
10%	RADA	5%	14.1	25.46 ± 1.16	24.08	**25.42 ± 1.14**	24.07
	RADA+	5%	2.6	**11.65 ± 8.20**	3.52	39.84 ± 11.75	24.37
	RADA	10%	16.7	23.25 ± 0.95	22.26	**23.36 ± 0.87**	22.47
	RADA+	10%	6.7	**11.24 ± 7.81**	3.63	36.03 ± 11.01	22.81

Table 2. Summary results of the performance evaluation of the RADA and RADA+ algorithms with real data sets

Data sets	% Labeled	Algorithm	T	Train Error	Train Min.	Test Error	Test Min.
Page Blocks	5%	RADA	30.9	7.75 ± 6.22	0.73	9.55 ± 5.86	2.89
		RADA+	16.2	**0.67 ± 1.04**	0.07	**3.24 ± 0.38**	2.91
	10%	RADA	22.6	8.50 ± 8.52	1.69	9.72 ± 8.41	3.03
		RADA+	22.4	**0.90 ± 1.17**	0.22	**3.48 ± 0.53**	3.08
	20%	RADA	20.5	8.57 ± 1.96	7.11	7.90 ± 1.83	6.61
		RADA+	29.3	**3.46 ± 4.35**	0.95	**4.93 ± 3.14**	2.96
Wave Forms	5%	RADA	9.2	**0.35 ± 1.34**	0.00	11.29 ± 0.44	10.23
		RADA+	24.5	0.40 ± 1.55	0.00	**10.73 ± 0.55**	9.99
	10%	RADA	2.5	**0.99 ± 2.17**	0.13	12.77 ± 0.79	10.53
		RADA+	19.3	1.00 ± 2.15	0.02	**11.21 ± 0.62**	10.30
	20%	RADA	5.4	3.39 ± 2.73	1.42	11.09 ± 0.43	10.05
		RADA+	30.0	**2.69 ± 3.09**	0.71	**10.26 ± 0.49**	9.62

We tested two real data sets: Page Blocks and Wave Forms. In these data sets we changed the number of classes, mainly because both data sets had more than two. Table 2 shows the summary results of the performance evaluation on these real data sets of the RADA and RADA+ algorithms.

We must note that as the training information decreases, the performance gap between the proposed algorithm RADA+ and RADA becomes larger. Note that the difference in the training error is quite noticeable. This is due to the use of the framework EM in the algorithm, specially when the labeled data is scarce, which is the same result obtained for the synthetic data sets. In the T column, we observe a different behavior regarding the results obtained for the synthetic data sets. The number of iterations is always for the RADA+ algorithm than for RADA, however the minimum test error is lower, wish indicates that RADA+ reaches a smaller error.

6 Concluding Remarks

The results were mixed, mainly because of the difference in the data sets used in experiments. In the real data sets, RADA+ outperforms RADA in both of the data sets, however the results obtained in the Page Blocks experiments were better than the ones obtained in Wave Forms. In the synthetic data set the performance of RADA+ was only slightly better than the one obtained with RADA, but still there was an improvement.

It is important to analyze the effect of the algorithm Support Vector Machine (SVM) in the proposal. The SVM is an algorithm widely used in classification tasks, unfortunately it has a bad performance in the presence of data outliers. Thus, the use of SVM in this proposal is beneficial for noiseless data, but for noisy data it is rather harmful. This behavior is observed in the synthetic data sets, where the results obtained where not as good as the ones obtained in the real data sets. This leaves open the opportunity to explore the use of different supervised algorithms with the proposed RADA+. It is also likely that the semi-supervised learning paradigm suffers from outliers. Since it tries to use distributional information from the unlabeled data, if the data contains outliers that discovered distributional information might be misleading. Further studies are needed to prove these conclusions.

This paper does not intend to corroborate the robustness properties of the algorithm RADA, but rather use the concepts of Semi-Supervised Learning to improve performance of the algorithm with large amounts of unlabeled data. The experimental results proved that the performance of the RADA+ algorithm is better than the one for RADA under those conditions.

References

1. Allende-Cid, H., Salas, R., Allende, H., Ñanculef, R.: Robust Alternating AdaBoost. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 427–436. Springer, Heidelberg (2007)
2. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)
3. Bennett, K., Demiriz, A.: Semi-Supervised Support Vector Machines. Advances in Neural Information Processing Systems, 368–374 (1999)
4. Bennett, K.P., Demiriz, A., Maclin, R.: Exploiting unlabeled data in ensemble methods. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 289–296 (2002)
5. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning. MIT Press, Cambridge (2006)
6. Alche-Buc, F.d., Grandvalet, Y., Ambroise, C.: Semi-Supervised MarginBoost. Advances in Neural Information Processing Systems 1, 553–560 (2002)
7. Duda, R., Hart, P., Stork, D.: Pattern classification. Wiley-Interscience, Hoboken (2000)
8. Grandvalet, Y., d'Alché-Buc, F., Ambroise, C.: Boosting Mixture Models for Semi-supervised Learning. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 41–48. Springer, Heidelberg (2001)
9. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal processing magazine 13(6), 47–60 (1996)
10. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: Seventh IEEE Workshop on Applications of Computer Vision, vol. 1, pp. 29–36 (2005)
11. Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, Chichester (1998)