A Fast Scheme for Feature Subset Selection to Avoid Overfitting in AdaBoost

Luigi Rosa
L.S. "Ettore Majorana", Via Frattini 11 10137, Turin, ITALY
Mobile +39 3207214179
Email luigi.rosa@tiscali.it
Web Site http://www.advancedsourcecode.com

***Abstract*—AdaBoost is a well known, effective technique for increasing the accuracy of learning algorithms. However, it has the potential to overfit the training set because its objective is to minimize error on the training set. We show that with the introduction of a scoring function and the random selection of training data it is possible to create a smaller set of feature vectors. The selection of this subset of weak classifiers helps boosting to reduce the generalization error and to avoid overfitting on both synthetic and real data.**

Keywords-AdaBoost, Classifier, Overfitting, Feature selection.

I. INTRODUCTION

Boosting algorithms are an important recent development in classification. These algorithms belong to a group of voting methods, for example [1]-[3], that produce a classifier as a linear combination of base or weak classifiers. While empirical studies show that boosting is one of the best off the shelf classification algorithms theoretical results don't give a complete explanation of their effectiveness. Breiman [4] showed that under some assumptions on the underlying distribution "population boosting" converges to the Bayes risk as the number of iterations goes to infinity. Since the population version assumes infinite sample size, this does not imply a similar result for AdaBoost, especially given results of Jiang [5], that there are examples when AdaBoost has prediction error asymptotically suboptimal at $t=\infty$ (t is the number of iterations). It has been shown that AdaBoost has the potential to overfit [6]-[7], although rarely with low noise data. However, it has a much higher potential to overfit in the presence of very noisy data [8]-[9]. At each iteration, AdaBoost focuses on classifying the misclassified instances. This might result in fitting the noise during training. In this paper, we use selection of a proper subset of weak classifiers to adjust the hypothesis of the AdaBoost algorithm to improve generalization, thereby alleviating overfitting and improving performance.

II. EXISTING ALGORITHMS

Several approaches have been proposed to avoid overfitting in AdaBoost algorithm [12]-[16]. Given a fixed amount of training data, there are at least six approaches to avoiding underfitting and overfitting (Fig. 1), and hence getting good generalization: model selection, jittering, early stopping, weight decay, bayesian learning, combining classifiers. The first five approaches are based on well-understood theory. The best way to avoid overfitting is to use lots of training data but in some circumstances this is not possible. If we want to select a subset of appropriate features from the total set of features with cardinality D, we have a choice between 2^D possibilities. If we deal with feature vectors with more than 100 components, the exhaustive search takes too much time [10]-[11]. Thus, we must find other ways to select a subset of features. There are mainly two strategies for avoiding the complete search: random based and deterministic greedy algorithms:

- Forward selection. This technique starts with an empty set and greedily adds the best of the remaining features to this set. This process is called stepwise, if only one feature is added in each step.
- Backward elimination. Here, we start with the full set containing all features. Then we greedily remove the most useless features from this set. This process is called stepwise, if only one feature is added in each step.
- Random mutation. This strategy starts with a randomly selected feature set and adds randomly selected features or removes them from the set.

Alternatively, feature selection methods can be divided into filters, wrappers and embedded approaches. For analyzing the relevance of feature subsets, filters use evaluation functions that are independent from the learning method, while wrappers evaluate the feature subsets in respect to the learning result. In embedded approaches, the feature subset selection and the learning method are interleaved.

III. PROPOSED APPROACH

Let us assume that training set has M elements and let D the number of weak classifiers: from these classifiers we want to find a subset of P classifiers (with $P \leq D$) that is able to reduce generalization error. We randomly select N

elements (with $N \leq M$) from training set. This reduced training set has to include both negative and positive samples. On these data AdaBoost model is trained in a series of T rounds (with $P \leq T$). Final strong classifier will be a linear combination of T weak classifiers C_i with $1 \leq i \leq T$. To each weak classifier that has been selected a score is attributed. This procedure is repeated several times, and each time the training subset is randomly changed. At the end only P classifiers will be chosen with the top P highest scores. Several score functions $s(i)$ have been tested, but the scoring rule with highest performance is the linear one $s(i) = T - i$. Our final goal is the selection of best features to reduce the dimension of the feature space and to eliminate redundant features, improving generalization error.

IV. RESULTS

We have tested this scheme for feature subset selection for melanoma recognition [17]. The initial set of features has a cardinality D equal to 180. These weak classifiers have been developed using ABCD rule [18]-[19]. Image dataset was composed by 95 benign nevi (positive samples) and 25 malignant melanoma (negative samples), hence $M = 120$. Using a leave-one-out strategy with 94 positive samples and 24 negative sample we have obtained a recognition rate equal to 78.50% with AdaBoost. Then we tried to reduce the set of weak classifiers to $P = 22$. Several times we have randomly selected $N = 118$ samples and we have trained system with these data using $T = 24$ rounds, updating each time scores. At the end of this process we have selected classifiers with top P highest scores. After such reduction the recognition rate with AdaBoost has been increased up to 86.10% using leave-one-out cross validation.

V. CONCLUSIONS

We have provided a fast scheme to feature subset selection to avoid overfitting. The proposed algorithm becomes particularly effective when training set has a small cardinality. For greater image dataset the random selection of positive and negative samples helps the feature reduction algorithm to avoid that local minima are reached by AdaBoost even if training error converges to zero.

REFERENCES

[1] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, 55(1):119–139, 1997.

[2] L. Breiman, "Bagging predictors," Machine Learning, 24(2):123–140, 1996.

[3] L. Breiman, "Arcing classifiers," The Annals of Statistics, 26(3):801–849, 1998.

[4] L. Breiman, "Some infinite theory for predictor ensembles," Technical Report 579, Department of Statistics, University of California, Berkeley, 2000.

[5] W. Jiang, "On weak base hypotheses and their implications for boosting regression and classification," The Annals of Statistics, 30:51–73, 2002.

[6] A. J. Groves, and D. Schuurmans, "Boosting in the limit: Maximizing the margin of learned ensembles," Proc. Fifteenth Int. Conf. on Artificial Intelligence, 692–699, 1988.

[7] W. Jiang, "Process consistency for adaboost," Technical report, Department of Statistics, Northwestern University, 2000.

[8] T. Dietterich, "Ensemble methods in machine learning," Lecture Notes in Computer Science 1857:1–15, 2000.

[9] G. Ratsch, T. Onoda and K. R. Muller, "Soft margins for adaboost," J. of Machine Learning 42(3):287–320, 2001.

[10] M. Drauschke, "Feature Subset Selection with Adaboost and ADTboost," Technical Report, Department of Photogrammetry, University of Bonn, 2008.

[11] M. Drauschke and W. Förstner, "Comparison of Adaboost and ADTboost for Feature Subset Selection," PRIS 2008, pp. 113-122, 2008.

[12] T. Bylander and L. Tate, "Using Validation Sets to Avoid Overfitting in AdaBoost," in Proc. FLAIRS Conference, pp.544-549, 2006.

[13] Y. Sun, S. Todorovic, J. Li, "Reducing the overfitting of AdaBoost by controlling its data distribution skewness," Int. J. of Pattern Recognition and Artificial Intelligence (IJPRAI), vol. 20, no. 7, pp. 1093-1116, 2006.

[14] *G. Ratsch, T. Onoda and K. R. Muller, "An improvement of AdaBoost to avoid overfitting," in Proceedings of the 5th International Conference on Neural Information Processing (ICONIP'1998), 1998.*

[15] A. Vezhnevets, O. Barinova, "Avoiding Boosting Overfitting by Removing Confusing Samples," ECML 2007: 430-441, 2007.

[16] A. Vezhnevets, V. Vezhnevets "Modest AdaBoost - Teaching AdaBoost to Generalize Better," Graphicon-2005, Novosibirsk Akademgorodok, Russia, 2005.

[17] L. Rosa, "Automated Melanoma Recognition," http://www.advancedsourcecode.com/melanomarec.asp , 2011.

[18] E. Zagrouba and W. Barhoumi, "An accelerated system for melanoma diagnosis based on subset feature selection," Journal of Computing and Information Technology, 13(1), 69-82, 2005.

[19] E. Zagrouba, and W. Barhoumi, "A prelimary approach for the automated recognition of malignant melanoma," Image Analysis and Stereology Journal, 23(2), 121-135, 2004.

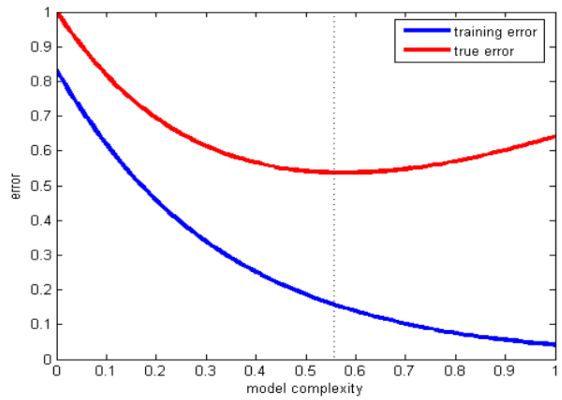

Figure 1. Training error and generalization error.