LEARNING OBJECT FROM SMALL AND IMBALANCED DATASET WITH BOOST-BFKO

Liansheng Zhuang[1], Wei Zhou[1], Qi Tian[2], Nenghai Yu[1]

[1] MOE-Microsoft Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China, Hefei, 230027, P.R.China
[2] Department of Computer Science, University of Texas at San Antonio, Texas, USA
Email: {ynh,lszhuang}@ustc.edu.cn

ABSTRACT

One of the main drawbacks of boosting is its overfitting and poor predictive accuracy when the training dataset is small and imbalanced. In this paper, we introduce a novel learning algorithm Boost-BFKO, which combines boosting and data generation. It is suitable for small and imbalanced training datasets. To enlarge training sets, Boost-BFKO uses the adaptive Balanced Feature Knockout procedure (BFKO) to generate new synthetic samples. To enrich the training sets, Boost-BFKO selects seed samples from the minority class, and rebalances the total weights of the different classes in the updated training dataset. Experiments on Caltech 101 database showed that our method achieves a desirable performance when only a few training samples are available for binary classification and multiple object classification.

Index Terms— gentleBoost, object detection, small training set，imbalanced data set

1. INTRODUCTION

Boosting is one of the most popular learning algorithms in computer vision [1]. It is widely used in object detection, such as face [2] and pedestrian [3]. It combines a group of weak classifiers, which perform slightly better than random guessing, into a strong classifier using weighted voting. Each weak classifier is based on a single feature component, and can be viewed as degenerate decision trees with a single node. During boosting training procedure, we search over all possible features. For each one, we form a classifier, and compute the total weight of misclassified samples. Then we select the one with the least weighted squared error as the weak classifier, and add it into previous classifiers forming a strong classifier. At each step, boosting reweights all the examples, and focuses on hard examples which are difficult to learn. After sufficient rounds, the strong classifier will reach the desired accuracy. If the dataset is large enough, boosting algorithms have low generalization properties. Otherwise, these algorithms tend to overfit and perform much worse than Support Vector Machine (SVM). If the dataset is imbalanced, these algorithms tend to bias towards the majority class, and produce classifier with poor predictive accuracy over the minority class [5]. In many real-world applications, it is difficult to collect a large number of training images. The number of positive samples is much fewer than that of negative. That is, the training set is often small and imbalanced.

To avoid overfitting, many methods have been proposed to learn from small samples [4, 6-8]. Recently, Wolf *et al.* has proposed a Feature Knockout (KO) procedure to generate new examples [4]. He applied KO to gentleBoost, and proposed the gentleBoostKO algorithm, which could work well with only a few examples. At each boosting round, gentleBoostKO fits a regression functions to each feature in the training set, and selects the regression function with the least weighted squared error as the weak classifier. The feature associated with the weak classifier is used for the KO procedure. In the KO procedure, an example (seed) is picked at random. Then, KO replaces one random feature value with the value of the corresponding feature of another example picked at random, and generates a new example. The new example is assigned with the same weight and class label of the seed. At the end of each round, the weights of all examples are updated and normalized. Though the number of training data is enlarged, the imbalance is preserved because of the random picking during the KO procedure. Assume that the output of KO was labeled as +1 or -1 for positive and negative with probability p and $(1-p)$ respectively. Here p is the percentage of positive examples. For example, if p =10%, the probability of generating a positive example will be only 0.1. Most of the new examples will be negative. This will aggravate the imbalance of training data.

Inspired by Wolf [4], in this paper, we propose an adaptive Balanced Feature Knockout procedure (BFKO) to enlarge and rebalance training datasets. Compared with original KO, our contributions include: (1) Instead of picking at random, we select seeds from the minority class at each round. This promotes the dataset balance, because most new examples belong to the minority class. (2) We replace all values of the features, which are associated with previous weak classifiers, with the probability $p_{replace}$. In

this way, many new examples are generated with similar values to their seeds. Adding these examples into training sets will improve the robustness of learned classifiers [4].

By imbedding BFKO into the gentleBoost framework, we propose a novel learning algorithm, Boost-BFKO, which is able to learn from small and imbalanced datasets. Experiments show that classifiers learned by Boost-BFKO have high predictive accuracy and robustness.

The rest of this paper is organized as follows. Section 2 describes the Boost-BFKO algorithm. A comparative evaluation on Caltech 101 database is followed in Section 3. Finally, we conclude the paper in Section 4.

2. BOOST-BFKO

Our BFKO procedure is illustrated in Algorithm 1. The major drawback of KO procedure is that it picks seeds at random, which causes that most of new synthetic examples belong to the majority class. This will aggrandize the imbalance of the training data. To avoid this, we hope that new examples will tend to bias towards the minority class, so that both classes are well represented during training. For this purpose, we select seeds from the minority class, and generate synthetic examples in the minority class space. The BFKO procedure does not change the corresponding thresholds. All the new synthetic examples have similar feature values to their seeds. This will cause the learned function smooth, and result in a more robust learned classifier according to noise injection theory [4].

Algorithm 1: The BFKO procedure

1. Select a seed set $\{(x_i, y_i)\}_{i=1}^{m}$ from the minority class. Here y_i is the label +1 or -1 of the minority class. m is the selected number and equals to $per_{\min} * N_{\min}$. $per_{\min}$ is the selected percentage and $N_{\min}$ is the total number of the examples of the minority class.
2. Select x_i and x_j from the seed set.
3. Replace each selected feature value of x_i with the corresponding value of x_j with the probability $p_{replace}$, generating new synthetic example $\hat{x}$. Here, we call x_i is the source of $\hat{x}$.
4. Repeat step 2 and step 3 to get different synthetic examples. All the synthetic are assigned to the same label as the minority class.

Because our feature knockout procedure generates new synthetic examples for the minority class, we call it *balanced Feature Knockout* procedure (BFKO). By using BFKO, we can rapidly enlarge the training samples to avoid boosting overfitting. Moreover, we can prevent the learning algorithm's bias towards the majority class by balancing the training dataset, and learn a robust classifier with more predictive accuracy. It is important for many real-world applications, because there are only a few samples available for training.

We imbed the BFKO procedure into the framework of gentleBoost, and propose a novel learning algorithm, Boost-BFKO, which is shown in Algorithm 2. Boost-BFKO is suitable to small datasets, as well as SVM, and can learn from imbalanced datasets.

Algorithm 2: The Boost-BFKO algorithm

Input: (x_1, y_1) ,…, (x_N, y_N) where $x_i \in R^n$, $y_i \in \{+1, -1\}$, and $i = 1,2,...,N$.

Output: Composite classifier $H(x)$.

1. Initialize weights $w_i \leftarrow 1/N$.
2. For $t = 1,2,...,T$
   (1) For each feature k, form a classifier function $f_t^{(k)}(x)$ by weighted least squares on y_i to x_i with weight w_i, $i = 1,...,N$.
   (2) Select the function $f_t^{(k_{\min})}(x)$ with the minimal associated weighted least square error as our weak classifier at this round.
   (3) Update the strong classifier $H(x)$:
   $$H(x) \leftarrow H(x) + f_t^{(k_{\min})}(x)$$
   (4) Select the seed set $S = \{(x_i, y_i)\}_{i=1}^{m}$ from the minority class with the percentage $per_{\min}$.
   (5) Create new synthetic examples $S' = \{(x'_j, y_j)\}_{j=1}^{m'}$ with the probability $p_{replace}$.
   (6) Set new example weight w'_j to that of its source: $w'_j \leftarrow w_{sourc}$.
   (7) Add new examples to training data set, and update the total number: $N \leftarrow N + m'$.
   (8) Update the weights and normalize:
   $$w_i \leftarrow w_i e^{-y_i f_t^{(k_{\min})}(x_i)}, \; i = 1,...,N$$
   $$w_i \leftarrow w_i / \sum_{i=1}^{N} w_i$$
3. Output the final classifier $H(x)$.

The high performance of Boost-BFKO mainly roots in its seeds selection at the same classes, instead of random selection. Random selection will increase the number of training examples of the majority class, aggrandizing the

imbalance of training set. On the other hand, random selection will generate a new synthetic example from two different classes at the same time. This will cause the new synthetic example far away from its source, and become an outlier. This is validated in our experiments in Section 3.

3. EXPERIMENT

In this session, we report our experimental results on Caltech 101 database [9]. In our experiments, we selected six categories, plus one background category as our training and test datasets. The training dataset included Cars side, Airplanes, Faces, Watch, Ketch, Motorbikes. Some sample images are shown in Fig. 1.

Fig.1 Some examples from Caltech 101 dataset [9].

We applied our Boost-BFKO to binary visual object classification, and compared the performance with that of SVM, gentleBoost and gentleBoostKO. In each experiment, we differed the images containing an object from the background images that do not contain the object. All experiments were repeated 8 times with different randomly selected training and test images. Similar to Wolf *et al.* [4], we also used the error at equilibrium-point as our error-measure. The result, shown in Fig. 2, was the average of 10 runs. For each run, we randomly selected 30 background images as our negative samples, and chose 3, 6, 10, 15 or 30 object images from each category as our positive samples. The rest images were the test dataset.

In our experiment, we chose the C2 feature as our image feature, and turned all the images into 500 C2 feature-vectors [10]. The C2 feature was inspired by biology and seemed extremely successful for learning to recognize objects [11]. We used the matlab code provided by Jim Mutch to extract our C2 feature [12].

In Fig. 2, it is clear that, when there are only a few examples for training, our method remarkably outperforms gentleBoost and gentleBoostKO. Compared with SVM, Boost-BFKO has the same performance level on the average for all the categories.

However, though SVM can be successfully applied to all the dataset from small to very large, the computational complexity of SVM at run time is costly. When classifying a new example, SVM needs computing all the features. If the feature number is large, it will spend much time to compute these features. This is unaffordable in some case, especially for multiclass object detection in clutter scenes. Boost-BFKO uses boosting techniques over weak classifiers based on single features, and the number of features used is easily controlled. In most case, the features used in the final strong classifiers are far less than those used by SVM, which results in a much faster classifier at run time.

Our future work include extending current binary classification to multi-class classification by enlarging the training dataset based on Genetic Algorithm and apply it to face and pedestrian detection.

4. CONCLUSION

In this paper, we propose an adaptive Feature Kockout procedure (BFKO), to enlarge and balance the training data set. It selects the seeds from the minority class to generate new synthetic examples. Then, it rebalances the total weights of all the updated training data set. We imbed the BFKO procedure into the framework of gentleBoost, and get a novel learning algorithm, Boost-BFKO, which is able to learn from small and imbalanced datasets. Experiments showed that classifiers learned by Boost-BFKO achieved a performance level as high as SVM but being more efficient, and outperformed gentleBoost and gentleBoostKO by a significant margin.

ACKNOWLEDGEMENTS

This paper is supported in part by National Natural Science Foundation of China (60672056), Specialized Research Fund for the Doctoral Program of Higher Education (20070358040), and ARO Grant W911NF-05-1-0404.

REFERENCES

[1]. J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Technical report*, Dept. of Statistics, Stanford University, 1998.

[2]. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. CVPR*, 2001, pages:I-511-I-518.

[3]. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. CVPR*, 2005, pages.886-893.

[4]. L. Wolf and I. Martin, "Robust boosting for learning from few examples," *Proc. CVPR*, 2005, pages:359-364.

[5]. M.A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, 2003.

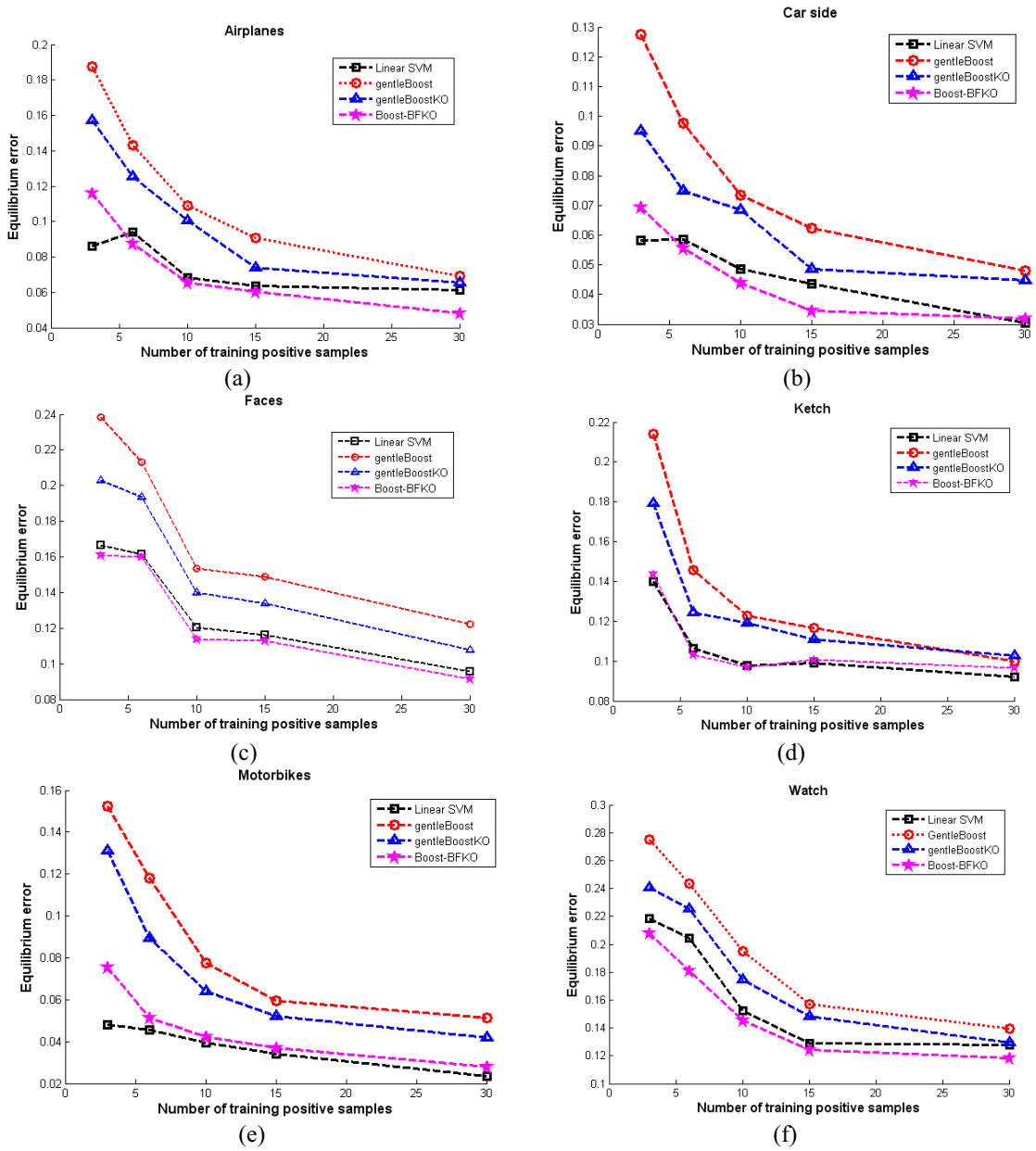

Fig.2 A performance comparison between Linear SVM, gentleBoost, gentleBoostKO and Boost-BFKO. (a) - (f) respectively show the performance on Airplanes, Car sides, Faces, Ketch, Motorbikes and Watch. The results are the average of 8 runs.

[6]. H. Guo, H.L. Viktor, "Learning from Imbalanced Data Sets with boosting and data generation: the DataBoost-IM approach," *ACM SIGKDD Explorations Newsletter,* 2004, 6(1), pages: 30-39.

[7]. A. Vezhnevets, O. Barinova, "Avoiding boosting overfitting by removing confusing samples," *Conf. ECML*, 2007, pages: 430-441.

[8]. J. Chen, R. Wang, S. Shan, X. Chen, and W. Gao, "Enhancing human face detection by resampling examples through manifolds," *IEEE Trans. On Systems, Man, and Cybernetics – Part A: Systems and Humans*, Nov. 2007, pages: 1017-1028.

[9]. F. F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004, pages: 178-178.

[10]. J. Mutch and D.G. Lowe, "Multiclass object recognition with sparse, localized features," *Proc. CVPR*, 2006, pages: 11-18.

[11]. F. F. Li, R. Fergus, and P. Perona, "A Bayesian approach to unsupervised one-Shot learning of Object categories," *Proc. ICCV*, Oc. 2003, pages: 1134-1141.

[12]. Jim Mutch Home Page. http://www.mit.edu/~jmutch