ON PREDICTING RARE CLASSES WITH SVM ENSEMBLES IN SCENE CLASSIFICATION

Rong Yan, Yan Liu, Rong Jin, Alex Hauptmann

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{yanrong, yanliu, rong, alex+}@cs.cmu.edu

ABSTRACT

Scene classification is an important technique to infer high-level semantic scene categories from low-level visual features. However, in the real world the positive data for many scenes may be rare, which degrades the performance of many classifiers. In this paper, we propose SVM ensembles to address the rare class problem. Various classifier combination strategies are investigated, including majority voting, sum rule, neural network gater and hierarchical SVMs. We also compare our method with two other common approaches for dealing with the rare class problem. Our experimental results show that hierarchical SVMs can achieve significantly better and more stable performance than other strategies, as well as high computational efficiency.

1. INTRODUCTION

Scene classification, which classifies images into meaningful semantic scenes by computing low-level visual features, has emerged as one of the important challenges in computer vision and machine learning. Over the past few years, some work to recognize high-level scenes has been done, such as indoor / outdoor classification [1, 2], cityscape / landscape classification [3] and hierarchical outdoor scene classification [4]. Their success demonstrates that the high-level scene properties can be inferred from the low-level visual features. However, most of this work is evaluated on balanced image datasets, in which the number of positive examples is comparable to that of the negative examples. Unfortunately, many general real-world image datasets only contain small numbers of positive examples for scene classification. For example, there are only less than 8% cityscape images and 3% landscape images in the training set of the TREC02 Video Track Feature Extraction Task [5]. One explanation for this is that the positive example of a scene is typically a coherent subset (e.g. Cityscape, Landscape, and Sunrise) of all the possible images, but the negative class is less well-defined as "everything else". Many of the learning algorithms will get in trouble when faced with imbalanced dataset [6], which limits the practical application of scene classification. Therefore, it is of crucial importance to study the problem of classifying rare classes in the scene classification.

To date, there have been a few attempts at addressing the real-world rare class problems as diverse as fraud detection [7], network intrusion, text categorization and web mining [8]. Some previous work has applied an ensemble-based approach, which is to combine several individual classifiers in some way to classify the test examples. In [7], a multi-classifier meta-learning approach has been devised to deal with skewed class distributions. More recent work [8] provides insight into when boosting, a strong ensemble-based learning algorithm, can achieve better precision and recall in the context of rare classes. This work claimed that the performance of boosting for rare class is critically dependent on the abilities of base learners.

In this paper, we propose a different ensemble-based approach that applies support vector machine (SVM) [9] ensembles to address the issue of predicting rare classes in scene classification. The idea of SVM ensembles is not new. Recent theoretical research [9, 10] has proposed SVM ensembles to adapt binary SVM to multi-class classification and address the high computational cost for training. However, our approach is different from previous study in several ways. First, the primary purpose differs considerably in that we applied SVM ensembles to address the rare class problem. Second, a different sampling scheme has been used and various combination strategies have been investigated in our paper. The experimental results demonstrate the high effectiveness and high efficiency of our approach.

The rest of the paper is organized as follows. Section 2 gives a brief overview of SVMs. Section 3 analyzes the effect of modifying the training distribution in the context of rare classes. Section 4 presents our approach of SVM ensembles and Section 5 presents the experimental results. Section 6 concludes the paper with a summary.

2. SUPPORT VECTOR MACHINES

SVMs [9] have been proposed with sound theoretical justifications to provide a good generalization performance compared to other algorithms [11]. The problem is to find a decision surface that "best" separates the data points into two classes based on the Structural Risk Minimization Principle. The decision function is of the form

$$y = sign\left(\sum_{i=1}^{N} y_i \alpha_i K(x, x_i) + b\right) \quad (1)$$

where x is the d-dimensional vector of a test example, $y \in \{-1, 1\}$ is a class label, x_i is the vector for the ith training example, N is the number of training examples, $K(x, x_i)$ is a kernel function, $\alpha = \{\alpha_1, ..., \alpha_N\}$ and b are the parameters of the model. These α_i can be learned by solving following quadratic programming (QP) problem,

$$\min Q(\alpha) = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x, x_i) \quad (2)$$

subject to $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C, \forall i$

The kernel function can have different forms, such as the polynomial kernel $K(u, v) = (u \cdot v + 1)^p$ and the Radial Basis Function (RBF) $K(u, v) = exp(-\gamma \|u - v\|^2)$ kernel . In our experiments, we choose the SVM with RBF kernel as the base classifier. [1]

3. EFFECT OF TRAINING DISTRIBUTION

In this section, we analyze the effect on the prediction performance of varying the distribution of negative/positive examples in the training set. Although SVMs are relatively insensitive to the distribution of training examples in each class, they will still get stuck when the class distribution is too skewed. The reason is that SVMs tend to generate the trivial model by almost always predicting the majority class, which is obviously not the desired classification result. [12] shows that the imbalance of datasets does degrade the prediction accuracy especially for non-linearly separable data. Although it is still an open question whether artificially varying the training distribution can improve the prediction performance for a rare class [6], [13] provides some insight and qualitative analysis of how tuning the distribution of the training set can help performance.
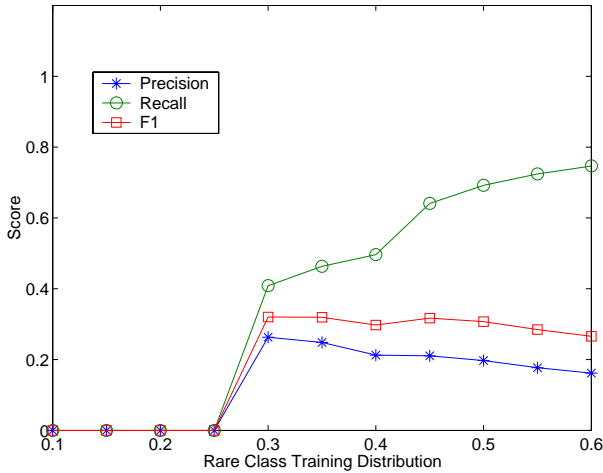

Fig. 1. The effect of modifying training distribution by over-sampling

Two basic methods were proposed for dealing with the rare class problem by modifying the class distribution [13]: Over-sampling, which is to replicate the data in the minority class, and under-sampling, which is to throw away part of the data in the majority class. Both methods mitigate the issue of imbalance in the dataset. However, both of them have known drawbacks. Under-sampling may eliminate some of potentially useful data, and the performance of classifiers suffers. Over-sampling, on the other hand, increases the training set size and the training time. This is a more critical problem to SVMs than other standard classifiers since the training time complexity for SVM is close to quadratic to the number of support vectors, even cubic in the worse case [14]. In addition, overfitting is more likely to happen with replication of minor examples [13].

To demonstrate the effect of varying the class distribution, we apply over-sampling for the scene classification data using a single SVM, altering the minority-class distribution from 10% - 60%. Figure 1 shows the performance for the cityscape dataset with respect to precision, recall and F1-measure [2]. By examining Figure 1, we observe that SVMs always predict the test examples all as negative and thus yields zero precision/recall until the size of the rare class examples is roughly comparable to the size of the majority class examples. This result again suggests varying the class distribution could improve the classification performance.

4. SVM ENSEMBLES

To address the drawback of over-sampling and under-sampling, we introduce the SVM ensembles to tackle the rare class problem. In this section, we will discuss the two main issues of our ensemble approach, i.e. the overall architecture and the combination strategies.
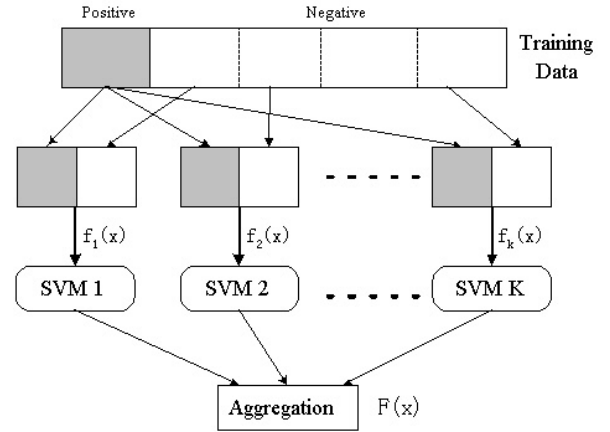

Fig. 2. Architecture of the SVM ensembles

4.1. Overall Architecture

As was shown in section 3, varying the training distribution could help the prediction of rare classes. However, we would like to generate the training sets with the desired distribution without either removing any training data or significantly increasing the training time. To achieve this, we use following strategies to sample the training sets. First, decompose the negative examples into K partitions, where K is depending on the number of positive examples, and combine all the positive examples with each partition of negative examples to be an individual subset. The next step is to train SVMs independently on every subset of the training set, and finally combine all constituent SVMs by various strategies explained in section 4.2. The architecture of our approach is depicted in Figure 2.

SVM ensembles has several advantages over the basic methods. First, SVM ensembles use the information of the entire dataset compared to under-sampling which uses only part of the dataset; in addition they reduce the computational cost dramatically compared to over-sampling. Second, SVM ensembles are able to overcome the limitation of the current SVM software. SVMs are stable

---

[1]Other kernels, such as the linear and polynomial kernel, have been tried and RBF kernel provides the best performance.

[2]The detailed definitions of F1-measure and summary of the cityscape dataset are presented in Section 5

classification methods and expected to learn exact parameters for a global optimum [11]. However, most current SVM software is implemented with approximation algorithm [14], and we cannot guarantee that a single SVM, which is used for the over/under-sampling case, always provides the optimal performance over the test data. SVM ensembles can overcome this limitation by combining those potentially suboptimal solutions and thus achieve better performance.

4.2. Combination of multiple SVMs

After each classifier is trained independently, we have to aggregate their results in an appropriate combination approach. How to combine the results of each classifier is an interesting research issue concerning SVM ensembles. Some combination strategies are suggested by previous studies: if only class labels are considered, a majority vote can be used; if continuous-valued outputs like posteriori probabilities are available, the sum of all the output probabilities has been suggested (Sum Rule) [15]. Besides these direct combination strategies, it is possible to "stacked" another learning method on top using the outputs of the input classifiers as new features [15, 16]. A mixture model of SVMs with another neural network (NN) as a gater has been proposed to solve very large-scale classification problem [17].

Apart from these combination strategies, we also propose hierarchal SVMs to address the rare class problem by using another SVM to aggregate the output of several SVMs. Formally speaking, let K be the number of decomposed training sets, let $f_k(k = 1...K)$ be the decision function of individual SVM training on training set and F be the final decision function of upper-layer SVM. The upper layer SVM is trained on a held-out set, which is sampled from the training set. The final decision $f_{SVM}(x)$ for a test vector x is determined by $f_{SVM}(x) = F(f_1(x), ..., f_K(x))$.

We are motivated to choose SVM as the top-level classifier for several reasons. First, the SVM-based combination can learn the combination weights automatically, while majority voting and probability-based combination treats all the classifiers with equal weights, even though the classifiers are not equally useful. Besides, SVM combination encourages local experts, and is relatively insensitive to the poor performed classifiers. Second, SVM has better generalization ability and requires less effort on tuning the parameters compared with neural network, which tends to have more stable performance.

One of the important issues for rare class prediction is that the training distribution is more likely to be different from the test distribution [6]. To address this, we need to sample a held-out set within the local region of the test set. In our approach, the held-out set is sampled by picking the nearest neighbor of each test example from the training set, where the distance is measured by cosine similarity. This sampling scheme is expected to have a better estimation of test set distribution than random sampling.

5. EXPERIMENTAL RESULTS

In this section, we demonstrate the efficiency and effectiveness of SVM ensembles with the training set of TREC02 Video Track Feature Extraction Task [5], which consists of approximately 23 hours of video. They cover a wide range of topics, such as natural scenes, man-made objects, cartoon and so forth. The images are extracted from the video every half second and manually labelled respectively as cityscape / no cityscape, landscape / no landscape. Over

100,000 images were labelled in our database. For each scene classification task, we randomly sample the image data to be the final dataset in order to reduce the computational cost. Table 1 provides the summary of the datasets.

In our experiment, we use SVM^{Light} [14] to train all the SVMs, running on a PIII1GHz with 512MB of RAM. The base classifier is the RBF-kernel SVM with parameter 0.05. Ten-folder cross validation was used for all our results. Some previous work suggested when predicting rare class problem, the F-measure is a more appropriate measure than other common metrics [18]. Therefore we adopt F1-measure as our major performance metric, which is defined as $F1 = 2PR/(P + R)$, where P, R is the precision and recall respectively We use the performance of the over-sampling and under-sampling strategies as our baseline because of their popularity in the literature.

TASK	POSITIVE	NEGATIVE	RATIO
Cityscape	190	2360	7.45%
Landscape	86	2760	3.02%

Table 1. Summary of the positive and negative examples for the cityscape and landscape datasets

5.1. Image Feature

Two kinds of low-level features are used in our experiment: color features and texture features. We generate these features for each subblock of a 3*3 image tessellation. The color feature is the central and second-order color moments for each separate color channel, where the three channels come from HSV color space. We use 16 bins for hue and 6 bins for both saturation and value. The texture features are obtained from the convolution of the subblock with various Gabor Filters [19]. In our implementation, 6 angles are used and each filter output is quantized into 16 bins. We compute a histogram for each filter and again generate their central and second-order moments as the texture feature. In total, we obtained 18 features for each subblock and concatenate them into a longer vector of 144 features for every image.

5.2. Classification Results

For each classification method, we vary the training distribution so that the rare class examples account for the following eleven distributions of each training set: 10%, 15%...55%, 60%. Six different classification methods are examined in Table 5: Over Sampling (OverSamp), Under Sampling (UnderSamp), SVM ensembles with majority voting (SVM-MV), SVM ensembles with Sum Rule (SVM-Sum), SVM ensembles with a neural network gater which has between 10 and 200 hidden units (SVM-NN) and hierarchical SVMs where the top-level SVM is linear kernel (SVM-SVM). For each method, we report the training distribution that achieves its best F1 performance and the corresponding results. We use the output value of SVMs directly as input features to the SVM ensembles except for SVM-Sum, in which the output value is scaled to a range between 0 and 1 as a posterior probability by logistic regression.

As can be seen from Table 5, over-sampling always achieves better performance than under-sampling at the price of higher computation cost. As a good alternative, hierarchical SVMs almost always outperform the other methods. Compared with over-sampling, hierarchical SVMs yield an 11% improvement on the cityscape dataset, and a 5% improvement on the landscape dataset in terms

	Cityscape					Landscape				
	Best Dist.	Prec	Rec	F1	Time	Best Dist.	Prec	Rec	F1	Time
OverSamp	30%	0.263	0.408	0.320	329.5	35%	0.225	0.679	0.339	454.1
UnderSamp	35%	0.237	0.397	0.297	26.44	35%	0.193	0.524	0.283	11.84
SVM-MV	35%	0.270	0.378	0.315	145.4	30%	0.305	0.286	0.295	150.1
SVM-Sum	45%	0.220	0.553	0.315	174.9	35%	0.271	0.308	0.288	151.2
SVM-NN	50%	0.263	0.277	0.270	185.1	50%	0.424	0.310	**0.358**	211.2
SVM-SVM	40%	0.326	0.387	**0.354**	150.1	40%	0.265	0.539	0.356	147.4

Table 2. Classification Results Using Different Methods For Tackling The Rare Class Problem. For Each Method, The Best Training Distribution For Rare Class (Best Dist.), Precision (Prec), Recall (Rec), F1 Metric (F1) And Training Time in Seconds (Time) are reported.

of F1 measure. In addition, it is much faster to train hierarchical SVMs, which only takes 1/3 - 1/2 training time of over-sampling. The performance of SVM-MV and SVM-Sum lies between over-sampling and under-sampling. We conjecture the reason for their low performance is their assumption of equal weight given to each classifier. SVM-NN produces quite unstable results, i.e. best in the landscape data set but worse in the cityscape, which indicates that neural network combination is sensitive to subtle characteristics of different datasets. Therefore, hierarchical SVMs are preferred to address the rare class problem taking both performance and computation cost into account.

6. CONCLUSION

In this paper, we have shown how to address the rare class problem in a framework of SVM ensembles for scene classification. We construct individual training sets by combining a subset of negative data with all the positive data, and aggregate the output value of each classifier. Various combination strategies are investigated in the TREC02 video track training dataset and hierarchical SVMs are found to yield better and more stable performance than other strategies, as well as lower computational cost than over-sampling. We conclude that hierarchical SVMs are good candidates to address the rare class problem in scene classification, which can simultaneously achieve high effectiveness as well as high efficiency.

7. REFERENCES

[1] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *IEEE International Workshop in Content-Based Access to Image and Video Databases, Bombay, India*, Jan 2002.

[2] A. Savakis N. Serrano and J. Luo, "A computationally efficient approach to indoor/outdoor scene classification," in *International Conference on Pattern Recognition, Qubec City, Canada*, Aug. 2002.

[3] A. Jain A. Vailaya and H.J. Zhang, "On image classification: City vs. landscape," in *IEEE Workshop on Content-Based Access of Image and Video Libraries, Santa Barbara, CA*, Jun 1998.

[4] A. Jain A. Vailaya, M. Figueiredo and H. Zhang, "A bayesian framework for semantic classification of outdoor vacation images," in *SPIE Conference on Electronic Imaging, San Jose, California*, 1999.

[5] TREC-2002 Video Track, "http://www-nlpir.nist.gov/projects/t2002v/t2002v.html," .

[6] Foster Provost, "Machine learning from imbalanced data sets 101/1," in *AAAI Workshop on Learning from Imbalanced Data Sets. Tech Rep. WS-00-05, Menlo Park, CA: AAAI Press*, 2000.

[7] P. Chan and S. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection," in *Proc. Fourth Intl. Conf. Knowledge Discovery and Data Mining*, 1998, pp. 164–168.

[8] V. Kumar M.V. Joshi, R.C. Agarwal, "Predicting rare classes: Can boosting make any weak learner strong?," in *the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada*, July 2002.

[9] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[10] E. Chang K. Goh and K. T. Cheng, "Svm binary classifier ensembles for multi-class image classification," in *ACM International Conference on Information and Knowledgment Management (CIKM), Atlanta*, November 2001.

[11] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 955–974, 1998.

[12] N. Japkowicz, "Learning from imbalanced data sets: a comparison of various strategies," in *AAAI Workshop on Learning from Imbalanced Data Sets. Tech Rep. WS-00-05, Menlo Park, CA: AAAI Press*, 2000.

[13] G.M. Weiss and F. Provost, "The effect of class distribution on classifier learning," Tech. Rep., Department of Computer Science, Rutgers University, 2001.

[14] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), Springer*, 1995.

[15] R.P.W. Duin J. Kittler, M. Hater and J. Mates, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, 1998.

[16] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, 1992.

[17] S. Bengio R. Collobert and Y. Bengio, "A parallel mixture of svms for very large scale problems," *Neural Computation*, vol. 14, no. 5, 2002.

[18] M.V.Joshi, "On evaluating performance of classifiers for rare classes," in *the Second IEEE International Conference on Data Mining (ICDM'02), Japan*, 2002.

[19] B. Schiele and J. L. Crowley, "Object recognition using multidimensional receptive field histograms," in *ECCV*, 1996.