SEMANTIC CONCEPT DETECTION IN IMBALANCED DATASETS BASED ON DIFFERENT UNDER-SAMPLING STRATEGIES

Jinlin Guo[1], Colum Foley[1,2], Cathal Gurrin[1,2], and Songyang Lao[3]

[1]School of Computing, Dublin City University, Ireland.
[2]CLARITY, Dublin City University, Ireland
{jguo, cfoley, cgurrin}@computing.dcu.ie
[3]School of Information System and Management, NUDT, China
laosongyang@vip.sina.com

ABSTRACT

Semantic concept detection is a very useful technique for developing powerful retrieval or filtering systems for multimedia data. To date, the methods for concept detection have been converging on generic classification schemes. However, there is often imbalanced dataset or rare class problems in classification algorithms, which deteriorate the performance of many classifiers. In this paper, we adopt three "undersampling" strategies to handle this imbalanced dataset issue in a SVM classification framework and evaluate their performances on the TRECVid 2007 dataset and additional positive samples from TRECVid 2010 development set. Experimental results show that our well-designed "under-sampling" methods (method SAK) increase the performance of concept detection about 9.6% overall. In cases of extreme imbalance in the collection the proposed methods worsen the performance than a baseline sampling method (method SI), however in the majority of cases, our proposed methods increase the performance of concept detection substantially. We also conclude that method SAK is a promising solution to address the SVM classification with not extremely imbalanced datasets.

Index Terms— Imbalanced Dataset, Classification, SVM, Under-sampling, TRECVid

1. INTRODUCTION

Semantic concept detection, also known as high-level feature extraction, is a research topic of great interest as it provides an alternative solution to the major scientific problem for video retrieval: the semantic gap [1]. After many years of research, the current emphasis on concept detection is to utilise a more generalized semantic indexing by classification learning algorithms [2] rather than utilizing domain-specific cues (or

knowledge) within the multimedia data which correlate with semantic concepts [3].

Theoretically, semantic concept detection can be processed by any supervised learning algorithm. However, this is not always valid in the real-world scenario, because learning algorithms often assume the positive/negative data distribution is balanced but multimedia collections usually contain only a small fraction of positive examples for semantic concepts. For example, in the TRECVid 2007 [4] training set, the imbalance for 20 concepts is shown in Fig. 1. Especially, there are only 8 shots labeled as concept *U.S. flag*. This is because the positive examples of a semantic concept is typically a coherent subset of keyframes, but the negative class is less well-defined as "everything else" in the collection. Unfortunately, many learning algorithms will face difficulties because of this imbalance. For instance, when the class distribution is too skewed, SVMs will generate a trivial model by predicting everything to the majority class, even though SVMs have been shown to be relatively insensitive to the distribution of training examples. Japkowiczaz [5] shows that the data imbalance issue can significantly degrade the prediction performance especially when the training data are non-linearly separable. Therefore, it's of crucial importance to address the rare data problem in the context of detecting concepts.

There has been a little work to date addressing the classification problem with imbalanced datasets. Existing work can be divided into two categories. One is based on improvements of classification algorithms, which aims to make these algorithms applicable to classification with imbalanced datasets by introducing some solutions to eliminate the influence of imbalance [6, 7, 8, 9]. In [6], Joshi et al. provided insights into the cases when AdaBoost, a strong ensemble-based learning algorithm, can achieve better precision and recall in the context of rare classes. They claimed that the performance of AdaBoost for rare class is critically dependent on the learning abilities of the base classifiers. Yan et al. [7] proposed an ensemble approach that first partitions negative data into small groups, constructs multiple classifiers using posi-
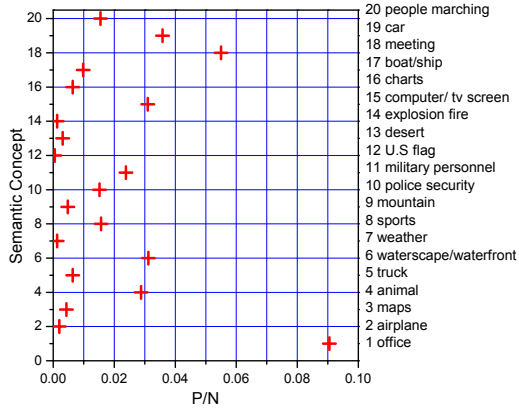
Fig. 1. Imbalance in TRECVid 2007 video dataset. *P* and *N* denote the size of positive and negative samples respectively. The y axis represents the concept No., and the x axis represents the proportion of positive samples to negative samples.

tive data as well as each group of negative data, and finally merges them via a top-level meta-classifier. Various classifier combination strategies are investigated including majority voting, sum rule, neural network and hierarchical SVMs. In [10], Dacheng Tao et al. analyzed the reason why small size of positive samples produce poor classification performance for SVM and they proposed asymmetric bagging and random subspace mechanism (with three algorithms) to address this biased classification problem.

As an alternative way to handle this problem. Approaches based on data sampling methods which aim to decrease the imbalance by increasing the size of positive class or reducing the size of the negative class artificially have been proposed in the literature [11, 12, 13]. Two of the most popular solutions are named "under-sampling" [11, 13], which throws away part of the negative data and "over-sampling" [11, 12] which replicates the positive data. In [11] Foster et al. provided insights and qualitative analysis of the effectiveness on tuning the training distribution. Over-sampling significantly increases the quantity of training data and thus consumes more time in the learning process. This problem is more critical to SVMs than other learning algorithms, since the training process for SVMs is very time-consuming. In [14], they reported that the training time complexity for SVMs is close to quadratic of the number of support vectors, even cubic in the worse case. In order to overcome the imbalance whilst decreasing the time for training, most of existing works only subsample randomly the domain class to obtain a roughly balanced training set [15].

In this paper, we employ three "under-sampling" strategies to handle SVM classification with imbalanced datasets for semantic concept detection. The first method is widely used which subsamples randomly in the majority class (in this paper, we select samples at intervals, named method SI). We propose two methods which select samples based on analyz-

ing the sample distribution in the majority class (method SNF and SAK), by which they aim to keep the information of the sample distribution. The proposed methods are evaluated on the TRECVid 2007 dataset, and additional positive samples from TRECVid 2010 development set. Experimental results demonstrate some promising results.

The rest of the paper is organized as follows. The strategies we adopt to balance the rare data are elaborated in section 2. Section 3 describes the experiment setup and presents the results. Section 4 concludes this paper.

2. BALANCING RARE DATA

In order to handle the rare class problem, three "under-sampling" strategies are adopted in the training stage of the SVM, namely method SI, SNF and SAK. In the following section, let N_v, N, P be negative samples set, the number of negative samples and the number of positive samples in the training data respectively. The flowchart for processing imbalanced data is show in Fig. 2.

Basic Biased SVM: Since the main problem existing in the classification with imbalanced data is that the classification hyper-plane is partial to majority class, which is inclined to mis-classify the rare class. The common solution to this problem is assigning different penalty parameters C^+ and C^- to the incorrectly classified positive and negative samples respectively. Therefore, the optimization problem is:

$$min \frac{1}{2} \left\| \omega^2 \right\| + C^+ \sum_{\{i|y_i=+1\}}^{l} \varepsilon_i + C^- \sum_{\{i|y_i=-1\}}^{l} \varepsilon_i$$

subject to:

$$y_i \left[(\omega x_i) + b \right] \geq 1 - \varepsilon_i \ (i = 1, 2, 3, ...l)$$
$$\varepsilon_i \geq 0 \ (l \ is \ the \ number \ of \ samples)$$

So that the primal formulation of the Lagrangian has two loss functions for the two types of errors:

$$L_p = min \frac{1}{2} \left\| \omega^2 \right\| + C^+ \sum_{\{i|y_i=+1\}}^{l} \varepsilon_i +$$
$$C^- \sum_{\{i|y_i=-1\}}^{l} \varepsilon_i - \sum_{i=1}^{l} \alpha_i \left[y_i \left(\omega \cdot \mathbf{x_i} \right) - 1 + \varepsilon_i \right]$$
$$- \sum_{i=1}^{l} \mu_i \varepsilon_i$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$. It's straightforward to show that the dual formulation gives the same Lagrangian as that of the primal SVM, but with the constrains as follows:

$$0 \leq \alpha_i \leq C^+, \ if \ y_i = +1$$
$$0 \leq \alpha_i \leq C^-, \ if \ y_i = -1$$

The α_i corresponding to the rare class with a non-zero slack variable is greater than that corresponding to the majority class with a non-zero slack variable. Therefore the classification hyper-plane is pushed towards the majority class.

Method SI: This intuitive method is commonly used in existing works and is our baseline, namely, a fixed number of the majority class is subsampled randomly to obtain a roughly balanced training set. In this paper, one is selected in every N/P negative samples, which produces the same size of negative samples as that of positive samples.
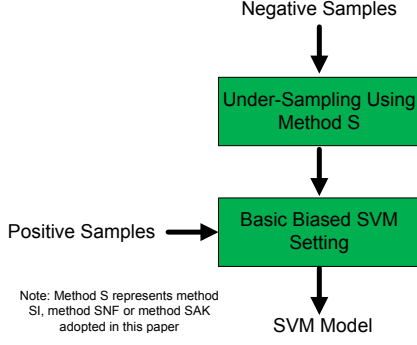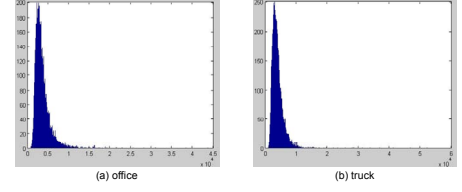
Fig. 2. Flowchart for processing imbalance in the training stage of the SVM classifier

Fig. 3. Examples of the distribution of distance from negative samples to the center of positive samples for 2 concepts

The next two methods are proposed based on such consideration: if we know the distribution of negative samples, then based on this distribution, selecting the similar size of positive samples may at least maintain the classification performance whilst decreasing the time requiring for training.

Method SNF: By analyzing positive and negative samples for 130 annotated semantic concepts in the TRECVid 2007 and 2010 datasets, we find that the Euclidean distance from one negative sample to the center C of positive samples follows approximately the Gaussian Distribution for each concept. Distributions for two concepts are shown in Fig. 3. Therefore, the training samples are selected according to the distribution of the Euclidean distance, which is described in detail by the following steps:

(1) Compute the center C of the positive samples and Euclidean Distance d_i from negative sample i to center C. D denotes the set of all the Euclidean distances and $D = \{d_1, d_2, d_3, ..., d_N\}$.

(2) The distribution curve of D is normally fitted as $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\mu)^2}{2\sigma^2}}$, where μ and σ are the fitted parameters, d is the random variable.

(3) Generating $P * \left(1 + \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\mu)^2}{2\sigma^2}} dx\right)$ random numbers from the Gaussian Distribution with parameters μ and σ. All the random numbers which are greater than zero are selected. The set of selected random numbers is denoted as R and $card(R) \approx P$

(4) For $\forall r \in R$, select the i_{th}^* negative sample as a training sample if $i^* = \underset{1 \leq i \leq N, i \notin T}{argmin} |d_i - r|$, where $d_i \notin D$ and T denotes the set of indexes of negative samples that have been selected as training data.

Method SAK: This method is based on such observation that if the negative samples can be divided into several clusters, then samples selected on these clusters quite likely represent the distribution of all negative data well. Based on

this consideration, we employ another under-sampling strategy named method SAK. The details are described as follows:

(1) Decompose the negative samples into K clusters. In our work, we use the k-means algorithm. Cluster i ($i = 1, 2, ..., K$) is denoted as $N_i = \{n_1^i, n_2^i, ..., n_{q_i}^i\}$, q_i denotes as the size of cluster i. And $\sum_{i=1}^{K} q_i = N$, $\bigcup_{i=1}^{K} N_i = N_v$. The cluster centers are denoted as $C = \{c_1, c_2, c_3, ..., c_K\}$. Set $i = 1$.

(2) For cluster i, compute $[q_i * P/N]$ cosine values, the m_{th} ($1 \leq m \leq [q_i * P/N]$) is computed as $o_m = cos(2\pi*N*m/(q_i*P))$. Compute the cosine-similarity between each negative sample n_j^i and its center c_i, namely, $s_{i,j} = Sim(c_i, n_j^i)$. Set $m = 1$.

(3) For the m_{th} cosine value o_m, select n_j^i as a training data, if $j = \underset{1 \leq j \leq q_i}{argmin} (|s_{i,j} - o_m|)$.

(4) $m+ = 1$. if $m \leq [q_i * P/N]$, repeat (3), else go to (5).

(5) i+=1, if $i \leq K$, repeat(2), else end.

After the under-sampling operation, an approximately equivalent number of negative samples and all the positive samples are chosen as the training data for the SVM classifier.

3. EXPERIMENT SETUP AND RESULTS

In this section, we compare the performances of these three under-sampling strategies on 20 semantic concepts based on the TRECVid 2007 dataset and additional positive samples from TRECVid 2010 development set.

3.1. Datasets

We test the performance of these three under-sampling strategies on 20 semantic concepts based on the TRECVid 2007 video dataset of news magazine, science news, news reports, documentaries, educational programs, and archival video in MPEG-1 format. About 50 hours are used to train the classifiers, which are segmented into shots. Each video shot is labeled with each of the 20 concepts by collaborative annotation. And 50 hours are used for evaluation purposes. The center frame is extracted as the keyframe for each shot. The

**Table 1**. Details of imbalance for 20 concepts in TRECVid 2007 dataset

Concepts	Positive #	Negative #
office(No.1)	1052	11621
airplane(No.2)	29	14797
maps(No.3)	64	14820
animal (No.4)	392	13597
truck (No.5)	90	14129
waterscape/waterfront(No.6)	408	13087
weather(No.7)	18	14175
sports (No.8)	220	13977
mountain (No.9)	69	14341
police security(No.10)	207	13584
military personnel (No.11)	328	13764
U.S flag(No.12)	8	15140
desert (No.13)	45	14497
explosion fire(No.14)	20	14846
computer/tv screen (No.15)	414	13357
charts(No.16)	90	13965
boat/ship (No.17)	136	13898
meeting (No.8)	710	12904
car (No.19)	478	13338
people marching(No.20)	221	14288

Table 2. Additional positive samples to six concepts

Concepts	Sizes
airplane(No.2)	163
maps(No.3)	186
weather(No.7)	26
desert(No.13)	101
explosion fire(No.14)	1194
U.S flag(No.12)	120

imbalance for positive and negative samples in the training data is shown in Fig. 1 and details are listed in Table 1. For detailed concept descriptions, please refer to the LSCOM [16].

The imbalance for the concepts *airplane* (No.2), *maps* (No.3), *weather* (No.7), *desert* (No.13), *explosion fire* (No.14) and *U.S flag* (No.12) is much more severe than that of other concepts, and both of method SNF and SAK achieve worse classification performance (cf. section 3.4.). Therefore, it's worth considering whether the classification performance will be better if more positive data are introduced. In order to test this, we introduce more positive samples to the training set. For each of these six concepts above, the corresponding positive samples from TRECVid 2010 development dataset are introduced. The size of additional positive samples in TRECVid 2010 development dataset for these concepts are listed in Table 2.
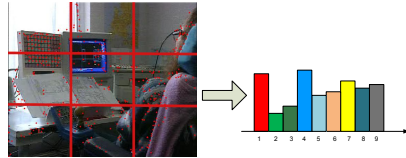
Fig. 4. Processing of SURF feature extraction

3.2. Implementation Details

For the SVMs, they are implemented using LIBSVM (Version 2.91) [17]. The RBF kernel is chosen for its good classification results comparing to polynomial and linear kernels [18].

The k-means algorithm used for clustering is sensitive to the choice of initial centers. Different initial centers may produce different result of clusters, and the algorithm may be trapped in the local optimum. In order to overcome this defect, we run k-means algorithm 20 times with different initial centers, and select the one with the least variance.

In our experiments, the penalty parameters C^+ and C^- for mis-classified positive and negative samples are set as $C^+ = 4$, $C^- = 1$ and the number of clusters is predefined as $K = 10$ based on our observations.

3.3. Low-level Features of Keyframes

Three MPEG-7 color and texture descriptors and one scale- and rotation-invariant descriptor SURF [19] are used as low-level features in our experiments. Three MPEG-7 descriptors are Color Layout (12 dimension), Scalable Color (64 dimension) and Edge Histogram (80 dimension). For the SURF feature, we adopt a histogram by grouping the interest points into regions. Given a keyframe and a set of keypoints, a 3×3 grid is defined. A 9-bin histogram which is a count of the keypoints that occur in each square is created (see Fig. 4). In total, all the features are concatenated into a vector of 165 dimensions for each keyframe.

3.4. Results and Analysis

For evaluation, we use the common measure from the TRECVID benchmarking: inferred average precision (infAP); infAP is similar to average precision (AP) in that it measures both precision and recall whilst taking into account rank position, but varies in that it makes use of sampled truth data, rather than complete truth data, More detail can be found in [20]. The experimental results are shown in Fig. 5.

As shown in Fig. 5, our first observation is for the concepts listed in Table 3, both method SNF and SAK worsen the infAPs when compared to method SI, about 8.9% (including *boat/ship* (No.17)) and 7.6% (including *people marching* (No.20)) respectively. When analyzing the results based on the imbalance in training data (see Fig. 1), we conjecture the reason for their lower performances is that the rare class prob-
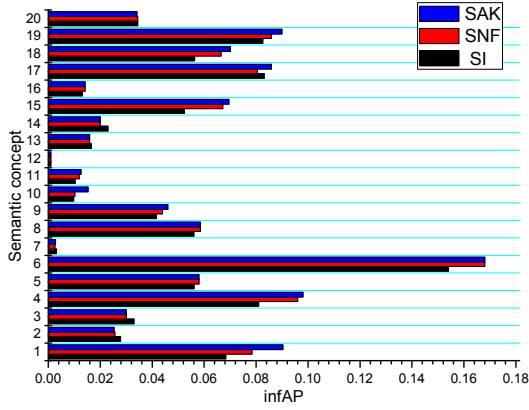
Fig. 5. Results Comparison of three strategies

Table 3. Method SI achieves better performance

Concepts	SI	SNF	SAK
airplane(No.2)	**0.0278**	0.0256	0.0253
maps(No.3)	**0.033**	0.03	0.03
weather(No.7)	**0.0031**	0.0026	0.0028
desert(No.13)	**0.0166**	0.0159	0.0159
explosion fire(No.14)	**0.023**	0.02	0.02
U.S flag(No.12)	**0.001**	**0.001**	**0.001**

Table 4. Method SNF and SAK get better infAPs

Concepts	SI	SNF	SAK
office(No.1)	0.0683	0.0784	**0.0903**
animal(No.4)	0.081	0.096	**0.098**
truck(No.5)	0.0561	**0.058**	**0.058**
waterscape/waterfronts(No.6)	0.154	**0.168**	**0.168**
sports(No.8)	0.056	**0.0585**	**0.0585**
mountain(No.9)	0.0416	0.044	**0.046**
police security(No.10)	0.0098	0.0102	**0.0153**
military personnel(No.11)	0.0103	0.012	**0.0126**
computer/tv screen(No.15)	0.0524	0.0672	**0.0695**
charts(No.16)	0.0131	**0.0142**	**0.0142**
boat/ship(No.17)	0.0831	0.0805	**0.0859**
meeting(No.18)	0.0563	0.0665	**0.0701**
car(No.19)	0.0826	0.0859	**0.0899**
people marching(No.20)	**0.0344**	**0.0344**	0.0341

Table 5. Results after introducing more positive samples

Concepts	SI	SNF	SAK
airplane(No.2)	0.031	0.028	**0.031**
maps(No.3)	0.031	**0.038**	0.035
weather(No.7)	**0.0033**	0.0026	0.0031
desert(No.13)	0.0171	0.021	**0.0241**
explosion fire(No.14)	0.042	0.046	**0.051**
U.S flag(No.12)	0.003	0.002	**0.005**

lem in the training set of these concepts is much more severe than that of other concepts and the negative samples selected by method SNF and SAK aren't capable of capturing the distribution of all the negative samples accurately.

However, for the 14 concepts listed in Table 4, method SNF and (or) SAK get better performance than method SI, increasing the performance by 9.7% (excluding *boat/ship* (No.17)) and 16.9% (excluding *people marching* (No.20)) respectively than SI. The reason for the increasing is possibly due to less imbalance in the training set for these concepts. For 14 concepts (including *U.S flag* (No.12) but *people marching* (No.20) excluded), method SAK gets the best results, especially for concept *office* (No.1), *police-security* (No.10) and *computer/tv screen* (No.15), the infAPs increase by more than 30% respectively. In total, for these 20 concept, method SNF and SAK increase the detection performance by 9.7% and 16.9% than SI respectively.

Furthermore, for concept *U.S flag* (No.12), three under-sampling strategies achieve the same and very low performance, only about 0.001 (as shown in Table 3), the reason for which may be attributed to two aspects. One is the discrimination power of the low-level features we select are very weak for this concept, and other more discriminative low-level features should be considered; The other is the imbalance of the training data is too prominent (just as aforementioned, $P/N \approx 0.0005$), no under-sampling approach is capable of capturing the distribution of all the negative samples

to some extent. Finally, for most of the concepts, method SNF achieves a median between the infAPs obtained by method SI and that by method SAK.

In order to test our hypothesis that the server imbalance causes the poor performance, we introduce more additional positive samples from the TRECVid 2010 development dataset for six concepts. Experimental results are shown in Table 5. Comparing Table 5 with Table 3, for nearly all the six concepts, the three under-sampling strategies achieve better performances after introducing more positive samples (except that for concept *maps* (No.3) obtained by method SI, and *weather* (No.7) by SNF), which is consistent with the common view that more annotated training set would gain performance. We could also find that the infAPs for most of the six concepts increase significantly, especially for concept *explosion fire* (No.14) and *U.S flag* (No.12,), although the infAP for concept *U.S flag* (No.12) is still very low, which may attribute to the low discrimination power of the low-level feature we used. Furthermore, except for concept *maps* (No.3) and *weather* (No.7), method SAK achieves the best performance among the three under-sampling strategies. In total, introducing more positive samples can boost the performance to some extent.

4. CONCLUSION

In this paper, we adopt three under-sampling strategies SI, SNF and SAK to address the imbalanced dataset problem in a SVM classification framework for semantic detection. Experimental results on TRECVid datasets show that, our well-designed "under-sampling" methods (method SNF and SAK) increase the performance of concept detection about 9.6% overall. In cases of extreme imbalance in the collection the proposed methods reduce the performance when compared to a simple baseline sampling method (method SI), however in the majority of cases, our proposed methods increase the performance of concept detection substantially. we also conclude that method SAK is a promising solution to address the SVM classification with not extremely imbalanced samples.

5. REFERENCES

[1] Arnold W. M. Smeulders, Senior Member, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, 2000.

[2] Cees G. M. Snoek, Marcel Worring, Jan C. Van Gemert, Jan mark Geusebroek, and Arnold W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM Multimedia*. 2006, pp. 421–430, ACM Press.

[3] David A. Sadlier, Noel O'Connor, Sean Marlow, and Noel Murphy, "A combined audio-visual contribution to event detection in field sports broadcast video. case study: Gaelic football," in *In Proceedings of ISSPIT'03*, 2003, pp. 14–17.

[4] P. Over, G. Awad, W. Kraaij, and A. Smeaton, "TRECVID 2007-overview," in *TRECVID 2007 Workshop Notebook Papers, Gaithersburg, MD, USA.*, 2007.

[5] Nathalie Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies," 2000, pp. 10–15, AAAI Press.

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

[7] Rong Yan, Yan Liu, Rong Jin, and Alex Hauptmann, "On predicting rare classes with svm ensembles in scene classification," in *In ICASSP*, 2003, pp. 21–24.

[8] Gang Wu and Edward Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *In ICML 2003 Workshop on Learning from Imbalanced Data Sets*, 2003, pp. 49–56.

[9] Kaizhu Huang, Haiqin Yang, Irwin King, and Michael R. Lyu, "Learning classifiers from imbalanced data based on biased minimax probability machine," in *Proc. CVPR*, 2004, pp. 558–563.

[10] Dacheng Tao, Xiaoou Tang, Xuelong Li, and Xindong Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, 2006.

[11] Gary Weiss and Foster Provost, "The effect of class distribution on classifier learning: An empirical study," Tech. Rep., 2001.

[12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[13] Miroslav Kubat and Stan Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *In Proceedings of the Fourteenth International Conference on Machine Learning*. 1997, pp. 179–186, Morgan Kaufmann.

[14] Thorsten Joachims, "Making large-scale support vector machine learning practical," 1998.

[15] Adrian Ulges, Christian Schulze, Markus Koch, and Thomas M. Breuel, "Learning automatic concept detectors from online video," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 429–438, 2010.

[16] Milind R. Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston H. Hsu, Lyndon S. Kennedy, Alexander G. Hauptmann, and Jon Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, 2006.

[17] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[18] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet, "Region-based video content indexing and retrieval," in *in CBMI 2005*, pp. 21–23.

[19] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

[20] Emine Yilmaz and Javed A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *CIKM*, 2006, pp. 102–111.