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 “under-sampling” 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. Japkowicz [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-
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 \( N_p, N, P \) 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:

\[
\min \frac{1}{2} \| \omega \|^2 + C^+ \sum_{i | y_i = +1} \varepsilon_i + C^- \sum_{i | y_i = -1} \varepsilon_i
\]

subject to:

\[
y_i (\omega^T x_i + b) \geq 1 - \varepsilon_i \quad (i = 1, 2, 3, \ldots l)\\
\varepsilon_i \geq 0 \quad (i \text{ is the number of samples})
\]

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

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

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:

\[
0 \leq \alpha_i \leq C^+, \quad \text{if } y_i = +1\\
0 \leq \alpha_i \leq C^-, \quad \text{if } y_i = -1
\]

The \( \alpha_i \) 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.
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 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, where $\mu$ and $\sigma$ are the fitted parameters, $d$ is the random variable.

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

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

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

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

(3) For the $m_{ih}$ cosine value $om$, select $n_{j}$ as a training data, if $j = \text{argmin}_{1 \leq j \leq q_i} |s_{i,j} - o_m|$. $m = m + 1$. If $m \leq [q_i \ast P/N]$, repeat (3), else go to (5).

(4) $i = i + 1$, 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

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

Table 2. Additional positive samples to six concepts

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplane (No.2)</td>
<td>163</td>
</tr>
<tr>
<td>maps (No.3)</td>
<td>186</td>
</tr>
<tr>
<td>weather (No.7)</td>
<td>26</td>
</tr>
<tr>
<td>desert (No.13)</td>
<td>101</td>
</tr>
<tr>
<td>explosion fire (No.14)</td>
<td>1194</td>
</tr>
<tr>
<td>U.S. flag (No.12)</td>
<td>120</td>
</tr>
</tbody>
</table>

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.

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 $3 \times 3$ 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-
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.16)) and 16.9% (excluding police security (No.11)) 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 undersampling 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.

![Fig. 5. Results Comparison of three strategies](image-url)
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