A Class Balanced Active Learning Scheme that Accounts for Minority Class Problems:
Applications to Histopathology

Scott Doyle¹, James Monaco¹, Michael Feldman², John Tomaszewski², and Anant Madabhushi¹

¹ Department of Biomedical Engineering, Rutgers University, USA, scottdo@eden.rutgers.edu, anantm@rci.rutgers.edu
² Department of Surgical Pathology, University of Pennsylvania, USA.

Abstract. Classifiers for detecting disease patterns in biomedical image data require manual annotations to serve as ground truth for training and evaluation, but are costly to obtain due to the complexity of the images and the expert medical knowledge required. An intelligent training strategy can maximize the efficiency of manual annotation. In this paper we present a novel class balanced active learning (CBAL) framework for classifier training to detect cancerous regions on prostate histopathology. The active learning (AL) algorithm identifies samples in a set of unlabeled data that will maximize the classification accuracy; only these samples are annotated, reducing the cost of training. We also address the minority class problem where one class (in this case, cancer) is under-represented. By using a query strategy that adds equal numbers of instances from both object classes (cancer and non-cancer) to the training set, each class is well-represented resulting in high classifier accuracy. Finally, we present a cost model of our CBAL strategy. We use the CBAL framework to train a classifier for finding cancer in images of prostate histopathology, and compare its accuracy against training strategies using random learning (RL) and those that do not enforce equal proportion of instances from both classes. On a dataset of over 12,000 prostate image regions, we find that (1) using CBAL the resultant classifier achieves the maximum possible accuracy (i.e. accuracy obtained by using all available samples for training) by using two orders of magnitude fewer samples, and (2) the predicted cost of CBAL agrees well with the empirically determined cost, which is not significantly higher than RL.

1 Introduction

Quantitative analysis of medical image data [1], [2] can greatly increase the ability of physicians to detect and diagnose disease states, but building ground

* Funding for this work made possible by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, ARRA-NCI-3 R21 CA127186-02S1, R21CA127186-01, R03CA128081-01), Department of Defense (W81XWH-08-1-0145), the Cancer Institute of New Jersey, and the Life Science Commercialization Award from Rutgers University.
Fig. 1. Annotation (black contour) of prostatic adenocarcinoma on digital histopathology. Cancer often appears near and around benign tissue, making annotation of cancer ground truth difficult and time-consuming.

trut for training and evaluation requires costly manual labeling. In particular, annotation of histopathology image data remains expensive [3], since: (1) Expert medical knowledge is required to perform accurate annotation. (2) Tissue digitized between 20-40x optical magnification (working pathology resolution) generates images several gigabytes in size, placing a large burden on the annotating pathologist. (3) The complex growth patterns of cancer, its proximity to benign tissue, and the presence of tissue types that mimic cancer make annotation a time-consuming process.

Figure 1 shows an image of prostate tissue, with cancer regions manually annotated via a black contour. Cancerous regions are intermingled with benign regions, complicating the annotation task and translating into several hours of a pathologist’s time. Additionally, the majority of the tissue area is non-cancerous, leading to the “minority class problem” where the target class (cancer) is under-represented compared to the non-target class. If this disparity were to be reflected in the training set, classification accuracy would suffer due to the lack of information about the target class. Thus, to minimize the cost of annotation and consequently to build an accurate classifier, we consider two issues: (1) which of the unlabeled samples should be annotated? and (2) what should the class balance of the final training set be? In this work, we focus only on the two class case, although extensions to the multi-class case should be relatively straightforward.

2 Active Learning and Class Equality

Active learning (AL) is a training strategy where only samples likely to improve classification accuracy are annotated. AL uses a classifier trained on a small subset of labeled data to find which of a pool of unlabeled samples are most likely to improve classifier performance, in contrast to drawing samples at random (i.e. random learning). AL methods are well-suited to large datasets that are costly to label, which is precisely the situation with annotation of histopathological data. Different AL methods apply various criteria for selecting samples: some methods [4, 5] employ distribution variance to choose samples, while others [6]
use the Query-by-Committee (QBC) approach, wherein disagreement among several weak binary classifiers is used to identify samples. These methods reduce the cost of building a training set by ensuring that each annotation increases classification accuracy, rather than randomly adding samples that may have little effect on overall classification accuracy. However, they do not explicitly consider the minority class problem, where the number of samples from one class is far less compared to the other.

The minority class problem is particularly prevalent in histological image analysis, where typically less than 10% of the total image is comprised of the target class (typically the disease class). Weiss, et. al [7] suggest that using a training set with an unbalanced class distribution may yield suboptimal classifier performance, since there is insufficient information available to accurately distinguish between the two classes. Additionally, the abundance of non-target samples introduces bias into the training procedure. Methods to adjust for the minority class include feature weighting [8], over-sampling minority samples [9], or enforcing an equal distribution of classes, which we refer to as “class balance.” Setting an a priori restriction on class distribution will increase the cost of annotation, since the pathologist will need to label enough samples to generate the desired training set makeup, but the tradeoff is increased accuracy.

In this work, we present a novel AL scheme for training a classifier that considers both (1) the problem of selecting informative samples, and (2) the minority class problem. When applied to the problem of analyzing regions of histological tissue images for disease, our methodology improves accuracy, sensitivity, and specificity over both random learning and training methods that do not account for the minority class. We also present a cost model for predicting the number of queries required to balance classes, and compare our model prediction to the empirical cost of the active and random learning strategies.

The rest of the paper is organized as follows. Section 3 explains the AL paradigm and class balancing. Experimental setup and results are given in Section 4, and concluding remarks in Section 5.

3 Minority Class Active Learning Algorithm

3.1 Bootstrap Training Set and Classifier Construction

The active learning algorithm ActiveTrainingStrategy is shown in Figure 3. The dataset is divided into an unlabeled training pool, $S^{tr}$ and an independent labeled testing pool, $S^{te}$ to test the accuracy at each step of the algorithm. The data comprises a set of square image regions $r \in R$, which are represented by the red squares in Figure 2 (a). Active learning is an iterative algorithm; we denote the training set at iteration $t \in \{0, 1, \cdots, T\}$ as $S^{tr}_{t,\Phi}$, where $\Phi$ denotes the training methodology and $T$ is the maximum number of iterations determined by a stopping criterion (Section 3.4). For $t = 0$, $S^{tr}_{0,\Phi}$ is randomly drawn from $S^{tr}$ ($S^{tr}_{0,\Phi} \subset S^{tr}$).

At iteration $t$, we construct a fuzzy classifier, $T_t(r) \in \{0, \frac{1}{M}, \frac{2}{M}, \cdots, 1\}$ trained on the current training set $S^{tr}_{t,\Phi}$ using the bagging algorithm developed
Fig. 2. An image (a) with regions $r$ outlined in red is classified using image features (b) and the confidence scene $R$ is constructed (c). An informative region (red, magnified in (d)) is chosen for annotation. The corresponding image section (e) is annotated; the lower left region (f) belongs to the cancer class.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^t, S^T$</td>
<td>Unlabeled training and testing pools</td>
<td>$T_t$</td>
<td>Fuzzy classifier using $S^T_{\phi}$</td>
</tr>
<tr>
<td>$S^E, S^E_{\phi}$</td>
<td>Eligible samples, annotated samples</td>
<td>$\Phi$</td>
<td>Training methodology</td>
</tr>
<tr>
<td>$r \in R$</td>
<td>Dataset of image patches</td>
<td>$S^t_{\phi}$</td>
<td>Samples labeled via $\Phi$ at $t$</td>
</tr>
<tr>
<td>$t \in {0, \cdots, T}$</td>
<td>Iteration of ActiveLearn</td>
<td>$k_1, k_2$</td>
<td>Total samples in $S^E$ from $\omega_1, \omega_2$</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of votes used to generate $T_t$</td>
<td>$\omega_1, \omega_2$</td>
<td>Possible classes of $r$</td>
</tr>
<tr>
<td>$R = (R, T_t)$</td>
<td>Confidence scene of regions $r \in R$</td>
<td>$k_1, k_2$</td>
<td>Total samples in $S^E$ from $\omega_1, \omega_2$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Classifier-dependent threshold for $T_t$</td>
<td>$N_t$</td>
<td>Samples added to training set at $t$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Stopping criterion threshold</td>
<td>$\tau$</td>
<td>Confidence margin</td>
</tr>
</tbody>
</table>

Table 1. Listing of the notation used in this paper.

by Breiman [10], where a set of $M$ weak binary classifiers are trained on a sub-sample of the training set. We are dealing here with the two-class case, although extensions to the multi-class case are possible. The classifier result is the average of these $M$ weak binary classifiers, where $T_t(r) = 1$ indicates a strong membership in class $\omega_1$, $T_t(r) = 0$ indicates strong membership in class $\omega_2$, and $T_t(r) = 0.5$ indicates an intermediate confidence. Shown in Figure 2 (c) is a confidence scene, $R = (R, T_t)$, where the intensity of a region $r \in R$ is given by the output of $T_t(r)$: bright regions correspond to regions with a high confidence in class $\omega_1$, dark regions correspond to a high confidence in class $\omega_2$, and gray regions represent intermediate confidence (uncertain).
**Algorithm: ActiveTrainingStrategy**

**Input:** $S^t$, $T$

**Output:** $T_T$, $S_T^t$

```
begin
0. initialize: create bootstrap training set $S^t_{0,φ}$, set $t = 0$
1. while $t < T$
2. Create classifier $T_t$ from training set $S^t_{0,φ}$;
3. Find eligible sample set $S^E$ where $T_t(r) = \frac{M}{2} \pm τ$;
4. Annotate $K$ eligible samples via MinClassQuery() to obtain $S^E$;
5. Remove $S^E$ from $S^t$ and add to $S^t_{t+1}$;
6. $t = t + 1$;
7. endwhile
8. return $T_T$, $S_T^t$;
end
```

**Fig. 3.** The active learning training process that uses MinClassQuery to obtain the samples for annotation, ensuring equal numbers of samples from class $ω_1$ and class $ω_2$.

### 3.2 Finding Informative Samples in the Unlabeled Pool

The fuzzy classifier is used to find a set of samples, denoted $S^E$, in the unlabeled training pool $S^t$ that are informative; that is, samples for which the confidence in classification is low. These are defined as samples for which:

$$T_t(r) = \frac{M}{2} \pm τ,$$

(1)

where $τ$ is the confidence margin indicating the selectivity of the algorithm. Smaller values of $τ$ define a smaller area on the interval $[0, 1]$, requiring more uncertainty for a region to be selected. The operating point of the fuzzy classifier is chosen as the point where there is an intermediate level of confidence in the classification of $r$; that is, where $T(r) = \frac{1}{2}$. The eligible samples described by Equation 1 are included in $S^E$. Algorithm MinClassQuery, shown in Figure 4, is used to query an expert for labels while enforcing an equal class distribution.

### 3.3 Cost Modeling of Class-Specific Querying

Algorithm MinClassQuery is employed to query $S^E$ until a specific class balance is achieved, and return the balanced set of annotated samples $S^E$. The number of samples $r \in S^E$ corresponding to classes $ω_1$, $ω_2$ are denoted $k_1$ and $k_2$, respectively (these values are unknown at the beginning of the querying process). The probability of randomly observing $ω_1$ in $S^E$ is denoted as

$$p_t(ω_1) = \frac{k_1}{k_1 + k_2},$$

(2)

referred to as the frequency-based estimate [7]. In the two-class scenario, $p_t(ω_2) = 1 - p_t(ω_1)$ and $p_t(ω_1) < p_t(ω_2)$ because $ω_1$ is in the minority class. These are parameterized by $t$ as the training pool changes at each iteration of the algorithm.
Algorithm: *MinClassQuery*

**Input:** $S^E$, $K > 0$, $\hat{k}_1$, $\hat{k}_2$

**Output:** $\hat{S}^E$

```
begin
0. initialize $\hat{S}^E = \emptyset$, $k'_1 = 0$, $k'_2 = 0$
1. while $|\hat{S}^E| \neq K$
2. Find class $\omega_i$ of a random sample $r \in S^E$, $i \in \{1, 2\}$;
3. if $k'_i < \hat{k}_i$
4. Remove $r$ from $S^E$ and add to $\hat{S}^E$;
5. $k'_i = k'_i + 1$;
6. else
7. Remove $r$ from $S^E$;
8. endif
9. return $\hat{S}^E$;
end
```

**Fig. 4.** Query strategy for obtaining new annotations while maintaining class balance.

We assume that the class distribution of the eligible samples is similar between the active and random learning cases. To enforce balanced classes we specify the number of samples in class $\omega_1$, denoted as $\hat{k}_1$, and from class $\omega_2$, denoted $\hat{k}_2$, to add to the training set. The total number of annotated samples is denoted $N_t$, and the number annotated in class $\omega_2$ is $N_t - \hat{k}_1$. Note that the total number of samples added to the training set, $K = \hat{k}_1 + \hat{k}_2$, is less than $N_t$. The probability of observing $k_1$ samples from class $\omega_1$ after annotating $N_t$ samples is given by the binomial distribution:

$$P_t = \sum_{k=0}^{\frac{N_t}{k_2}} [p_t(\omega_1)]^{k_1} [1 - p_t(\omega_1)]^{N_t-k_1}. \quad (3)$$

The cost of training depends on the value of $N_t$ required before $P_t \geq 0.5$ (i.e. when it is likely that $\hat{k}_1$ samples from class $\omega_1$ have been annotated). Because the number of annotations alters the summation in Equation 3, there is no closed-form solution for determining $N_t$ a priori. However, because we know $\hat{k}_1$ and we can estimate $p_t(\omega_1)$, we can solve the binomial cumulative distribution function (CDF) to find the $N_t$ for which $P_t \geq 0.5$ via any standard optimization or search strategy. The probabilities are then updated:

$$p_{t+1}(\omega_1) = \frac{k_1 - \hat{k}_1}{k_1 + k_2 - N_t}, \quad (4)$$

and $N_{t+1}$ is recalculated via the CDF. The cost of the entire querying procedure is calculated by summing $N_t$ for all $t$:

$$L = \sum_{t=1}^{T} N_t. \quad (5)$$
Thus we can calculate a priori the cost for any training task where $p_0(\omega_1)$, $\hat{k}_1$, and $\hat{k}_2$ are known. We can make a reasonable guess for $p_0(\omega_1)$ based on the size of the target class observed empirically ($< 10\%$). The total number of iterations $T$ corresponds to the size of the final training set and can be found using the stopping criterion discussed below. Classifiers that only perform well when given a large training set will require a large value for $T$, increasing cost.

### 3.4 Evaluation and Stopping Criterion

We obtain a hard classification for $r \in S$ as:

$$\hat{T}_t(r) = \begin{cases} 1 & \text{if } T_t(r) > \theta \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is a classifier-dependent threshold obtained via training. For region $r$, the ground truth label is denoted as $G(r) \in \{0, 1\}$, where a value of 1 indicates class $\omega_1$ and 0 indicates class $\omega_2$. Accuracy is calculated as:

$$A(T_t) = \frac{1}{|R|} \sum_r \begin{cases} 1 & \text{if } G(r) = \hat{T}_t(r) \\ 0 & \text{otherwise.} \end{cases}$$

The algorithm repeats until one of two conditions is met: (1) $S^v$ is empty, or (2) the maximum number of iterations $T$ is reached. A stopping criterion can be trained offline to determine the value of $T$ as the smallest $t$ that satisfies:

$$|A(T_t) - A(T_{t-1})| \leq \delta,$$

where $\delta$ is a similarity threshold. An assumption in using this stopping criterion is that adding samples to the training set will not decrease classifier accuracy.

### 4 Experimental Setup and Results

#### 4.1 Data Description and Feature Extraction

We apply the above training methodology to the problem of prostate cancer detection from biopsy samples. Glass slides containing prostate biopsy samples are digitized at 40x magnification and is divided into a set of square regions, $r \in R$, where $R$ is the set of all regions. Regions are 30 pixels square based on the size of the tissue structures that distinguish cancer. Ground truth annotation for cancer regions is performed manually by an expert pathologist. A total of 100 images were analyzed yielding over 12,000 image regions.

In [11] we have identified several hundred textural features capable of discriminating between cancerous and non-cancerous regions. Of these features, 14 highly discriminating features were selected that capture the texture differences between cancerous and benign tissue, including first-order statistical features, second-order co-occurrence features, and steerable Gabor wavelet features (Figure 2 (b)). These pixel-wise features are calculated for each 30-by-30 region and averaged to obtain a single feature value for that region.
4.2 Experiment List

We perform classification using five different training frameworks:

- **Class Balanced Active Learning (CBAL):** Methodology is as described above.
- **Unbalanced Active Learning (UBAL):** Does not consider class balance.
- **Class Balanced Random Learning (CBRL):** All unlabeled samples are queried, keeping class balance constant as described in MinClassQuery().
- **Unbalanced Random Learning (UBRL):** All unlabeled samples are queried randomly without regard for classes.
- **Full Training (FULL):** All available training samples are used, no query or active learning strategy.

In random learning (RL), all samples in the unlabeled pool $S_{tr}$ are “eligible” for querying; that is, $S_{re} = S_{tr}$. In unbalanced class experiments, the MinClassQuery algorithm is replaced by simply annotating $\hat{k}_1 + \hat{k}_2$ random samples (regardless of class) and adding them to $S_{re}$. The FULL training strategy represents the scenario when all possible training data is used (i.e. the highest possible accuracy). The classifier is tested against an independent testing pool, $S_{te}$. In these experiments, $T = 40$, the confidence margin was $\tau = 0.5$, and the number of samples added at each iteration was $K = 2$. In the balanced experiments, $\hat{k}_1 = \hat{k}_2 = 1$. A total of 12,588 image regions were used in the overall dataset; 1,346 were randomly selected for $S_{te}$, and 11,242 for $S_{tr}$. The true ratio of non-cancer to cancer regions in $S_{tr}$ was approximately 25:1 (4% belonged to the cancer class).

Two different classifier algorithms are employed to generate $T$: (1) decision trees created via the C4.5 algorithm [12]; and (2) support vector machines [13] which create a separating hyperplane in high-dimensional space.

4.3 Performance Measures for Evaluation

For evaluation, we classify the independent test set $S_{te}$ at each iteration, using accuracy defined in (7) and receiver operating characteristic (ROC) curves to determine the classifier’s ability to discriminate cancer from non-cancer regions. ROC curves compute the sensitivity and specificity of a fuzzy classifier by varying the threshold $\theta \in \{0, \cdots, 1\}$ to obtain a binary classification of the test set $S_{te}$. Each value of $\theta$ corresponding to a single point on the ROC curve, and the area under the curve (AUC) measures how well the classifier can discriminate between cancer and non-cancer regions.

4.4 Qualitative Results

Examples of confidence scenes $R$ are shown in Figure 5. Figures 5 (a), (d) show images with benign regions marked in red boundaries and cancerous regions in black. Figures 5 (b), (e) show the confidence scenes $R$ obtained via the CBAL training strategy, and (c), (f) are obtained via CBRL training. The intensity of the regions represents classifier confidence. Note that in both CBAL and
Fig. 5. Qualitative results of the final classifier. Shown are (a), (d) the segmented cancer region, (b), (e) the probability scene obtained through the AL classifier, and (c), (f) the probability scene obtained via random sampling.

CBRL, the cancer region (indicated by black boxes in Figures 5 (a), (d)) is identified as highly likely to be cancer. However, the RL images have more regions of uncertainty than the AL images, indicating that the AL classifier is more confident in the class of each region compared to the RL classifier.

4.5 Classifier Accuracy and Receiver Operating Curves

Quantitative classification results are plotted in Figure 6 as accuracy (Figures 6 (a), (c)) and area under the ROC curve (Figures 6 (b), (d)) as a function of the number of training samples in the set $S_t^r$ for $1 \leq t \leq 40$. The first row illustrates the results for the decision tree classifier, while the second row shows results for the support vector machine classifier. In each plot, the full training set corresponds to the straight black line, CBAL is the solid blue line, CBRL is a pink dotted line, UBAL is a black dashed line, and UBRL is a red dashed line.

The AUC values for CBAL approach the FULL training with 40 samples in the decision tree experiment and 20 in the support vector machine experiment, while CBRL, UBRL, and UBAL have lower AUC at those sample sizes. Accuracy for CBAL is much greater than other methods in the decision tree experiment for over 20 samples, while the support vector machine yields comparable accuracy to the other methods. For our dataset, CBAL works best when between 40 and 50 samples are used, at which point AUC and accuracy is very similar to the full training set. CBRL, UBRL, and UBAL do not perform as well in that window.
for the majority of our experiments, requiring a larger number of samples to match the accuracy and AUC of CBAL.

4.6 Cost Modeling of MinClassQuery

Figure 7 (a) shows the cost of CBAL and CBRL. The simulated cost is found by solving for $N_t$ in Equation 3 with an initial class probability of $p_0(\omega_1) = 0.04$ (based on the class distribution seen in practice) and $\hat{k}_1 = \hat{k}_2 = 5$. The value of $N_t$ is plotted as a function of $t$ for the simulation (solid black line) and the empirical cost of using the CBRL (pink dotted line) and CBAL (solid blue line) methods. The size of the training set, $t \ast (\hat{k}_1 + \hat{k}_2)$, yields a cost that can be found by integrating on the interval from $[0, t]$; for the 50-sample case, $t = 5$ which yields a cost of 330 samples for the simulation, 337 for the CBAL training, and 324 for CBRL. Although unbalanced training incurs a cost of 50 samples, the tradeoff is in the accuracy and AUC of the resulting training set. Additionally, the 337 training instances required for CBAL to achieve maximum possible classification accuracy (i.e. the FULL training strategy) requires an order of magnitude fewer samples ($|S^U| = 12,000$).

Also note that $N_t$ decreases as $t$ increases due to the changes in the class distribution of the unlabeled dataset as samples are annotated and removed. Figure 7 (b) plots the value of $p_t(\omega_1)$ as a function of $t$, with the solid black line representing the simulation, the dotted pink line represents CBRL, and the solid blue line is CBAL. Because there are proportionately more samples from the $\omega_1$ class as the algorithm iterates, the number of queries required to maintain class balance is reduced. Finally, the class distribution of the entire population of unlabeled samples (CBRL) is not significantly different from the distribution of the eligible samples $S^U$ found via ActiveTrainingStrategy (CBAL), which validates the assumption made when using the same cost model to predict the cost in both training scenarios.

5 Concluding Remarks

In this work we present a class balanced active learning (CBAL) paradigm for annotating histopathological training sets that accounts for the minority class problem, as well as a cost model for predicting the cost of building a training set with balanced classes. Active learning (AL) allows for the classifier to select samples that will have the greatest positive impact on accuracy, while enforcing class balance allows the classifier to properly identify the minority class. When analyzing digital prostate tissue samples for presence of cancer, the CBAL training method achieved accuracy and AUC values similar to those obtained with the full training set using fewer samples than the UBAL, CBRL, or UBRL methods. While balancing classes involves more queries, classification accuracy increases. We show that for the specific problem considered in this work, CBAL requires approximately 50 training samples to achieve comparable accuracy to the full
Fig. 6. Quantitative results of the classifier, $T_t$, for $t \in \{1, 2, \ldots, T'_t\}$. Shown are (a) accuracy and (b) AUC values for the decision tree classifier, and (c) accuracy and (d) AUC values for the support vector machine.

training set. The number of queries required to balance classes is not significantly different between the active and random learning strategies, as the class distribution of the unlabeled data is similar to the distribution of the eligible samples identified via AL; however, CBAL yields higher accuracy and AUC than CBRL for the same number of queries. For histopathological datasets (which are difficult and costly to annotate), CBAL provides a framework for efficiently generating training sets. Future work will involve extensions of our framework to the multi-class case, where relationships between multiple classes with different distributions must be taken into account.

References

Fig. 7. (a) Plot of queries $N_t$ required for class balance as a function of $t$; shown are CBAL (blue line), CBRL (red dashed line), and the value from Equation 3 (black line). (b) Plot of $p_t(\omega_1)$ as a function of $t$; the class distribution is not significantly different between the entire unlabeled pool of samples and the eligible samples identified by AL.