

# Statistical Learning for Effective Visual Information Retrieval

(Invited paper by IEEE International Conference on Image Processing, Barcelona, 2003)

Edward Y. Chang

VIMA Technologies & University of California, Santa Barbara

## Abstract

For effective retrieval of visual information, statistical learning plays a pivotal role. Statistical learning in such a context faces at least two major mathematical challenges: scarcity of training data, and imbalance of training classes. We present these challenges and outline our methods for addressing them: *active learning*, *recursive subspace co-training*, *adaptive dimensionality reduction*, *class-boundary alignment*, and *quasi-bagging*.

## 1 Overview

The principal design goal of a visual information retrieval system is to return data (images or video clips) that accurately match users' query concepts. To achieve this design goal, the system must first comprehend a user's query concept thoroughly, and then find data that match the concept in the low-level input space accurately. Statistical learning techniques can assist achieving the design goal via two complementary avenues: *semantic annotation* and *query-concept learning*.

*Semantic annotation* provides visual data with semantic labels to support keyword-based searches. Several studies propose semi-automatic annotation methods to propagate keywords from a small set of annotated images to the other images [2, 13]. Although semantic annotation can provide some relevant query results, annotation is often subjective and narrowly construed. When it is, query performance may be compromised. To thoroughly understand a query concept, with all of its semantics and subjectivity, a system must learn the target concept from the user directly via *query-concept learning* [10, 12]. (*Semantic annotation* can assist, but not replace, *query-concept learning*.)

At first sight, traditional supervised learning methods such as neural networks, decision trees, and Support Vector Machines (SVMs) can be directly applied to perform *semantic annotation* and *query-concept learning*. Unfortunately, "classical" learning algorithms are not adequate to deal with the technical challenges posed by these two tasks. To illustrate, let  $D$  denote the number of low-level features,  $N$  the number of training instances,  $N^+$  the number of positive training instances, and  $N^-$  the number of negative training instances ( $N = N^+ + N^-$ ). Two major technical challenges arise:

**1. Scarcity of training data.** The features-to-semantic mapping problem often comes up against the  $D > N$  challenge. For instance, in the query-concept learning scenario, the number of low-level features that characterize an image ( $D$ ) is greater than the number of images a user would be willing to label ( $N$ ) during a relevance feedback session. As pointed out by D. Donoho [3], the theories underlying "classical" data analysis are based on the assumption that  $D < N$  and  $N \rightarrow \infty$ , but when  $D > N$ , the basic methodology which was used in the classical situation is not strictly applicable.

**2. Imbalance of training classes.** The target class in the training pool is typically outnumbered by the non-target classes ( $N^- \gg N^+$ ). For instance, in a  $k$ -class classification problem where each class has about the same number of training instances, the target class is outnumbered by the non-target classes by a ratio of  $k - 1 : 1$ . The class boundary of imbalanced training classes tends to skew toward the target class when  $k$  is large. This skew makes class prediction less reliable.

In the following sections, we present our proposed methods for dealing with the above challenges. Section 2 presents the training-data scarcity problem in more detail and outlines our three proposed remedies: *active learning*, *recursive subspace co-training*, and *adaptive dimensionality reduction*. Section 3 presents the problem of training-class imbalance and outlines our two remedies, *class-boundary alignment* and *quasi-bagging*. In Section 4, we offer our concluding remarks.

## 2 Challenge I. Scarcity of Training Data

The major challenge that the *query-concept learning* task faces is the scarcity of training data. This is because users generally are too impatient to label a large number of training instances for the concept-learning task. A typical learning task might be to infer a target concept with fewer than one hundred training instances, although the number of low-level features (the dimension of the input space<sup>1</sup>) is usually much higher ( $D > N$ ).

*Semantic annotation* suffers from the training-data scarcity problem in a slightly different way. While it might be technically feasible to label a large number of training images, doing so would not be economically practical. Given

<sup>1</sup>The distinction between *input space* and *feature space* is important. Input space refers to the space where the feature vectors reside. Feature space refers to the space onto which the feature vectors are projected via kernel methods.

a fixed amount of budget and time to label  $N$  training instances, we would like the labeled pool to provide maximum information for learning a classifier that can propagate annotation with maximal accuracy [11].

We consider three approaches to address the training-data scarcity ( $D > N$ ) challenge. First, given  $N$  training instances from which to select, we choose the  $N$  most informative (most useful) instances for learning the target concept. Second, using the selected  $N$  instances, we infer additional training instances for increasing  $N$ . Third, dimensionality reduction methods can be employed for reducing  $D$ .

### Approach I.1. Making $N$ Instances Most Useful

The first challenge of *query-concept learning* is to find some relevant objects so that the concept boundary can be fuzzily identified. Finding a relevant object can be difficult if only a small fraction of the dataset satisfies the target concept. For instance, suppose the number of desired objects in a one-million-image dataset is 1,000 (0.1%). If we randomly select 20 objects per round for users to identify relevant objects, the probability of finding a positive sample after five rounds of random sampling is just 10%—clearly not acceptable.

We can improve the odds with an intelligent sampling method MEGA (the Maximizing Expected Generalization Algorithm) [8, 10], which finds relevant instances quickly, to initialize *query-concept learning*. MEGA models query concepts in  $k$ -CNF<sup>2</sup>, which can formulate virtually all practical query concepts. It uses  $k$ -DNF to bound the sampling space from which to select the most informative samples for soliciting user feedback. Even if all samples are labeled negative by the user, MEGA uses these negative-labeled instances (irrelevant instances) to shrink the sampling space (by making the  $k$ -DNF more specific). Thus we increase the probability of finding relevant instances in the next feedback iteration.

Once some relevant and some irrelevant instances have been identified, we can employ SVM<sub>Active</sub> [12] to refine the class boundary. Intuitively, SVM<sub>Active</sub> works by combining the following three ideas:

1. SVM<sub>Active</sub> regards the task of learning a target concept as one of learning an SVM binary classifier. An SVM captures the query concept by separating the relevant images from the irrelevant images with a hyperplane in a projected space, usually a very high-dimensional one. The projected points on one side of the hyperplane are considered relevant to the query concept and the rest irrelevant.
2. SVM<sub>Active</sub> learns the classifier quickly via active learning. The active part of SVM<sub>Active</sub> selects the most informative labeled instances with which to train the SVM classifier. This step ensures fast convergence to the query concept in a small number of feedback rounds.
3. Once the classifier is trained, SVM<sub>Active</sub> returns the top- $k$

<sup>2</sup>**Definition 1:**  $k$ -CNF: For constant  $k$ , the representation class  $k$ -CNF consists of Boolean formulae of the form  $c_1 \wedge \dots \wedge c_\theta$ , where each  $c_i$  is a disjunction of at most  $k$  literals over the Boolean variables  $x_1, \dots, x_n$ . No prior bound is placed on  $\theta$ .

most relevant images. These are the  $k$  images farthest from the hyperplane on the query concept side.

In short, we use MEGA to find initial relevant images, and then switch to SVM<sub>Active</sub> for refining the binary classifier and ranking returned images. In addition to using SVMs as the base classifier, we have also explored efficient implementation of BPMs and produced effective systems [2, 6].

### Approach I.2. Increasing $N$

As counter-intuitive as this may sound, the scarcity of negative training instances is more problematic than the scarcity of positive training instances for our learning algorithms. This is because describing a concept such as “tigers” may be challenging, but describing a negated concept (in this case the “non-tiger” concept) can require potentially infinite information. (The number of images needed to adequately portray the tiger concept is finite, but the number of non-tiger images is infinite.) Negative training instances are significantly easier to come by, but at the same time, the number of negative samples needed to depict a negated concept can be very large. Thus, we need a substantial number of negative training instances to accurately characterize the class boundary.

Let us revisit the SVM<sub>Active</sub> algorithm. At each round of sampling, it selects objects that are in the margin and also close to the dividing hyperplane. If the margin is wide (because of the lack of support at the negative side), the probability is low that a positive training instance can be found within that wide margin. Our goal is to find enough positive training instances to depict the target concept. However, the difficulty in gathering sufficient negative training instances slows down progress for finding positive training instances. The learning process is analogous to mining gold in a stream. The most productive way to harvest gold (i.e., relevant instances) is to quickly filter out water, mud and sand (i.e., negative-labeled instances).

For solving the problem of insufficient training data, *transductive learning* has been proposed to work with different learning algorithms for various applications. The basic idea of transduction is to leverage the unlabeled data near the labeled data. Suppose the nearest neighbors of negative training instances are always negative. We can then use the neighboring instances to substantially increase the negative-labeled pool. However, transduction must be performed with care, since the mis-labeling of data caused by wrong “guesses” can degrade the learning result.

The performance of transduction is, unfortunately, consistently inconsistent. Its performance could be not only application-dependent, but also dataset-dependent. For image retrieval, transduction may not be suitable for the following two reasons:

- Nearest-neighbor sampling. Most image query-concepts are not convex, nor do they reside continuously in the input space. For instance, flowers of different colors tend to spread out in the color-feature dimension. Thus, to describe either a flower or a non-flower concept, we might need to

explore the input space more aggressively. Transductive learning, however, exploits only in the nearest neighborhood of the existing labeled instances. Suppose a picture of a lake (one of many non-flower instances) is labeled negative. Using the pictures neighboring this lake picture, which are likely to contain lakes or something visually similar, may not be very productive in refining the description of the negated concept (non-flower).

- Noise factor. Mislabeled data due to noise can reduce the prediction accuracy of the classifier. In the norm case where relevant images are rare, performing transduction runs a very high risk of introducing false relevant instances, which thereby reduces class-prediction accuracy.

We propose employing *co-training* to increase  $N$  in a more meaningful and accurate way [5]. The problem of using a large unlabeled sample pool to boost performance of a learning algorithm is considered within the framework of co-training [1]. A broad definition of co-training is to provide each instance with multiple distinct views. Distinct views can be provided by different learning algorithms, e.g., by using MEGA and SVM<sub>Active</sub> at the same time. Information sharing between distinct views can increase especially the negative-labeled pool. Here, we propose another co-training method, which provides each training instance with distinct views via subspace learning to boost the pool of negative-labeled instances. As in the *query-concept learning* scenario, where the percentage of positive-labeled instances is rare, boosting the positive pool directly may not be productive. Thus, we attempt to boost the negative pool so that the probability of finding positive instances can be increased. This method recursively conducts subspace co-training at each feedback iteration in the following steps:

1. Divide the input space into  $G$  subspaces.
2. Conduct parallel training in these  $G$  subspaces using labeled training dataset  $L$ .
3. Use the  $G$  resulting learners to label the unlabeled pool and yield a new set of labeled instances  $L'$ .
4.  $L \leftarrow L \cup L'$
5. Go back to Step 1 until no more labeled instances can be inferred (i.e., until  $L' = \emptyset$ ).

### Approach I.3. Reducing $D$

When  $D$  is large, the training instances in the high-dimensional input space become very sparse (even when  $N > D$ ). The sparsely populated space causes two theoretical obstacles for statistical inference. First, given a query instance, its nearest neighbors tend to be equally distanced. Second, the nearest neighbors are not “local” to the query point. These two mathematical properties make class prediction difficult.

Many methods have been proposed for performing dimensionality reduction, e.g., principal component analysis (PCA) and independent component analysis (ICA). The major drawback of these methods in the context of *query-concept learning* is that they reduce dimensionality in a universal fashion

with respect to the entire dataset, and with respect to all users. Reducing dimensionality this way entirely disregards individuals’ subjectivity. Our work [9, 7] shows that a perceptual distance function is not only partial in dimensionality, but also *dynamic* in the subspace where the similarity between two images is measured. We thus propose using DPF (dynamic partial function) to perform adaptive dimensionality reduction.

A seamless way to integrate this partial and dynamic dimensionality reduction method into *query-concept learning* is to introduce a new DPF kernel function as the following:

$$K(x, x') = e^{-\sum_{j \in \Delta_m} |x_j - x'_j|^2 / 2\sigma^2}, \quad (1)$$

where  $\Delta_m$  is the set that contains  $m$  out of  $D$  dimensions that have the smallest distance.

## 3 Challenge II. Imbalance of Training Classes

A subtle but serious problem that hinders a classifier is the skewed class boundary caused by imbalanced training data. To illustrate this problem, we use a 2D checkerboard example. The checkerboard divides a  $200 \times 200$  square into four quadrants. The top-left and bottom-right quadrants are occupied by negative (majority) instances, but the top-right and bottom-left quadrants contain only positive (minority) instances. The lines between the classes are the “ideal” boundary that separates the two classes. In the rest of this section, we will use *positive* when referring to minority instances, and *negative* when referring to majority instances.

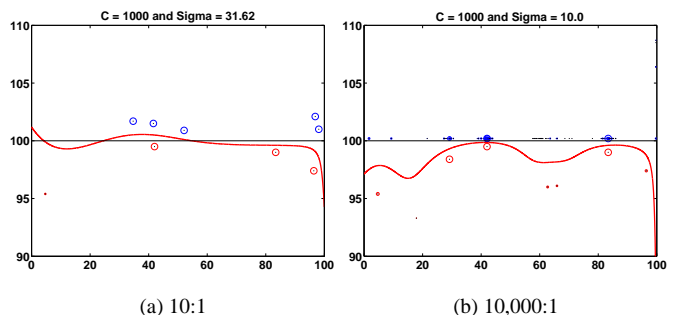


Figure 1: Boundaries of Different Ratios ( $N_n/N_p$ ).

Using SVMs as the classifier, we plot in Figures 1(a) and 1(b) the boundary distortion between the two left quadrants of the checkerboard under two different negative/positive training-data ratios. Figure 1(a) shows the SVM class boundary when the ratio of the number of negative instances (in the quadrant above) to the number of positive instances (in the quadrant below) is 10 : 1. Figure 1(b) shows the boundary when the ratio increases to 10,000 : 1. The boundary in Figure 1(b) is much more skewed toward the positive quadrant than the boundary in Figure 1(a), which results in a higher incidence of false negatives.

Both *query-concept learning* and *semantic annotation* may suffer from the imbalanced training-class problem, since the target class is always overwhelmingly outnumbered by

the other classes (that is, the relevant class is often rare and outnumbered by the irrelevant class). We consider two approaches for remedying the imbalanced training-data problem for SVMs: the algorithmic approach and the data-processing approach. The algorithmic approach focuses on modifying the kernel function or the kernel matrix, and the data-processing approach focuses on processing the training data.

### Approach II.1. Algorithmic Remedy

Examining the class prediction function of SVMs

$$\text{sgn}(f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b), \quad (2)$$

we can identify three parameters that affect the decision outcome:  $b$ ,  $\alpha_i$ , and  $K$ . In [14], we show that the only effective algorithmic method for tackling the problem of imbalanced training classes is through modifying the kernel  $K$  or the kernel matrix<sup>3</sup>  $K_{i,j}$  adaptively to the training data distribution. We propose using adaptive conformal transformation to dynamically enlarge (or scale) the spatial resolution of minority and majority areas around the separating boundary surface in the feature space, such that the SVM margin could be increased while the boundary becomes less skewed. Conformal transformation modifies the kernel function by changing the Riemannian geometrical structure induced by the mapping function. For data that do not have a vector-space representation (e.g., video sequence data), we apply transformation directly on the kernel matrix. Our experimental results on both UCI and real-world image/video datasets show the proposed algorithm to be very effective in correcting the skewed boundary. For further details, see [14].

### Approach II.2. Data Processing Remedy

Bagging subsamples training data into a number of *bags*, trains each bag, and aggregates the decisions of the bags to make final class predictions. We propose *quasi-bagging*, which performs subsampling differently than the bagging scheme does. For each bag, quasi-bagging subsamples only the majority class, but takes the entire minority class. By subsampling just the majority class, we improve class balance in each training bag. At the same time, using a large number of bags reduces the variance caused by the subsampling process. Let  $\hat{\theta}_n(\mathbf{x})$  denote a predictor function of a  $p$ -dimensional vector from data  $\mathbf{z} = \{(y_i, \mathbf{x}_i)\}_1^n$ . The quasi-bagging algorithm can be described as follows:

1. Randomly draw a bootstrap bag  $S_b \subset \{\mathbf{z}_i^*\}_1^m$  from the  $\mathbf{z}$ . The bag has  $m$  samples, including all minority examples.
2. Construct its bootstrap predictor  $\theta_{n,b}^*(\mathbf{x})$ .
3. Repeat Steps 1 and 2  $B$  times, and compute the aggregated predictor as  $\hat{\theta}_B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \theta_{n,b}^*(\mathbf{x})$ .

<sup>3</sup>Given a kernel function  $K$  and a set of instances  $S = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ , the kernel matrix (Gram matrix) is the matrix of all possible inner-products of pairs from  $S$ ,  $K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$ .

## 4 Conclusion

In this paper, we have shown that statistical learning plays a vital role in improving the effectiveness of visual information retrieval. We noted that “classical” learning methods are not suited for the task, so we presented several methods for tackling the challenges of training-data scarcity and training-class imbalance. Through extensive empirical studies [2, 10, 12, 14] and prototype implementation [6], these methods have been validated. Our future research will focus on designing indexing structures (e.g., [4]) that can support personalized visual information retrieval, and on devising kernel methods for conducting multi-camera security surveillance [15].

## References

- [1] A. Blum and T. Mitchell. *Combining Labeled and Unlabeled Data with Co-Training*. Proceedings of the Workshop on Computational Learning Theory, 1998.
- [2] E. Chang, K. Goh, G. Sychay, and G. Wu. Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. on Circuits and Systems for Video Technology Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description*, 13(1):26–38, 2003.
- [3] D. L. Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality. *American Math. Society Lecture — Math Challenges of the 21st Century*, 2000.
- [4] K. Goh, B. Li, and E. Chang. Dyndex: A dynamic and non-metric space indexer. *Proceedings of ACM International Conference on Multimedia*, pages 466–475, December 2002.
- [5] W.-C. Lai and E. Chang. Hybrid learning schemes for multimedia information retrieval. *IEEE Pacific Rim Conference on Multimedia*, pages 556–563, December 2002.
- [6] W.-C. Lai, E. Chang, K.-T. Cheng, and M. Crandell. PBIR-MM — multimodal image retrieval and annotation. *Proceedings of ACM Multimedia*, December 2002.
- [7] B. Li and E. Chang. Discovery of a perceptual distance function for measuring image similarity. *ACM Multimedia Journal Special Issue (to appear)*, 2003.
- [8] B. Li, E. Chang, and C.-S. Li. Learning image query concepts via intelligent sampling. *Proc. of IEEE Multimedia*, 2001.
- [9] B. Li, E. Chang, and C.-T. Wu. Dynamic partial function. *IEEE Conference in Image Processing*, September 2002.
- [10] B. Li, W.-C. Lai, E. Chang, and K.-T. Cheng. Mining image features for efficient query processing. *IEEE International Conference on Data Mining*, November 2001.
- [11] G. Sychay and E. Chang. Effective image annotation via active learning. *IEEE International Conf. on Multimedia*, 2002.
- [12] S. Tong and E. Chang. Support vector machine active learning for image retrieval. *Proceedings of ACM International Conference on Multimedia*, pages 107–118, October 2001.
- [13] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *Proc. of Interact: Conference on HCI*, pages 326–333, July 2001.
- [14] G. Wu and E. Chang. Adaptive feature-space conformal transformation for imbalanced-data learning. *International Conf. on Machine Learning*, August 2003.
- [15] G. Wu, Y. Wu, L. Jiao, Y.-F. Wang, and E. Chang. Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance. *UCSB Technical Report (submitted to ACM Multimedia)*, April 2003.