An Optimization Based Framework for Dynamic Batch Mode Active Learning

Shayok Chakraborty, Vineeth Balasubramanian and Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing (CUbiC)
Arizona State University
(schakr10, vineeth.nb, panch)@asu.edu

Abstract

Active learning techniques have gained popularity in reducing human effort to annotate data instances for inducing a classifier. When faced with large quantities of unlabeled data, such algorithms automatically select the salient and representative samples for manual annotation. Batch mode active learning schemes have been recently proposed to select a batch of data instances simultaneously, rather than updating the classifier after every single query. While numerical optimization strategies seem a natural choice to address this problem (by selecting a batch of points to ensure that a given objective criterion is optimized), many of the proposed approaches are based on greedy heuristics. Also, all the existing work on batch mode active learning assume that the batch size is given as an input to the problem. In this work, we propose a novel optimization based strategy to dynamically decide the batch size as well as the specific points to be queried, based on the particular data stream in question. Our results on the widely used VidTIMIT and the MBGC biometric datasets corroborate the efficacy of the framework to adaptively identify the batch size and the particular data points to be selected for manual annotation, in any batch mode active learning application.

1 Introduction

The fundamental goal in classification is to learn a function f which accurately maps data instances X into corresponding class labels y. To adequately learn a function with high generalization capability, it is imperative to acquire sufficient labeled data in the form of a training set. However, while gathering a huge amount of unlabeled data is cheap and easy, annotating large quantities of data (with class labels) is an expensive process in terms of time and human labor. This has paved the way for research in the field of *active learning*. Active learning algorithms automatically select the exemplar and representative instances, from a set of unlabeled data points, which are to be annotated manually. This tremendously reduces human labeling effort as only a few sample points, which are identified by the algorithm, need to be labeled manually. Several active learning algorithms have been proposed in the pattern recognition literature which can be categorized into 4 groups - (i) SVM based approaches [1], (ii) Statistical approaches [2], (iii) Ensemble based approaches [3] [4] and (iv) Other miscellaneous approaches [5][6].

All the aforementioned active learning algorithms query a single data instance at a time and update the classifier. However, the vast quantities of digital data that are being generated today necessitate a strategy to simultaneously select and learn from multiple data points. To address this need, *Batch mode active learning* (BMAL) schemes, which attempt to select a batch of points at one shot from the unlabeled set for manual annotation, have been proposed in recent years. Optimization based techniques can be judiciously used to handle this problem. Depending on the application, one can define an objective criterion and select a batch of points so as to optimize the value of the objective (e.g. minimize the variance of the future learner or maximize the log likelihood of the future learner with respect to the training set). However, most of the BMAL approaches that have been proposed

in literature are based on heuristic scores. Brinker [7] proposed a BMAL scheme which selected a highly diverse batch of points, where diversity was measured by the angle induced by the hyperplane of the selected point with all the other hyperplanes of the already selected points. Hoi *et al*. [8] used the Fischer information matrix as a measure of model uncertainty and proposed to select a batch of points which reduced the Fischer information of the classification model. The same authors applied the BMAL scheme to the problems of content based image retrieval (CBIR) [9] [10] and medical image classification [11]. Guo and Schuurmans [12] first formalized the problem by proposing an optimization based solution to select the most appropriate batch of unlabeled points for active learning.

In addition to the aforementioned examples of CBIR and medical image classification, BMAL algorithms are also highly relevant for applications involving video data. Modern video cameras have a high frame rate and consequently, the captured data has high redundancy. Selecting the promising instances from this superfluous set is a significant and valuable challenge. Due to its wide usage, we focus on face based biometric recognition systems as the exemplar application in this paper and explain the framework. Although validated only on biometric data in this work, the proposed framework is generic and can be used in any application where it is required to select a number of representative entities from repetitious samples.

All existing BMAL strategies require the batch size (the number of data points to be selected from an unlabeled set) to be specified in advance. In an application like face based biometric recognition, deciding on a batch size in advance and without any knowledge of the unlabeled video being analyzed, is impractical. The batch size should rather depend on the quality and variability of the data in the unlabeled video and also on the level of confidence of the current classifier on the images in the unlabeled stream. Similarly, the choice of the unlabeled data points for manual annotation requires careful specification of an objective function that is suitable for a given application. In this paper, we propose a novel optimization framework to simultaneously address two specific issues - (i) adaptively choose the batch size for a given set of unlabeled points and (ii) design an appropriate objective function that selects unlabeled data points that maximize the performance of the updated classifier, also ensuring that data points from low-density regions are selected. While we have designed an objective function for a video-based biometrics application, this function can be suitably tailored to the needs of other applications. Our work is motivated by the method proposed by Guo and Schuurmans [12], which however has a different objective and is restricted to static scenarios where the batch size is user specified. With the same computational complexity as [12], we solve for both the batch size as well as the specific points to be selected for annotation for this problem.

2 Problem Formulation

2.1 Batch Mode Active Learning for Biometrics

Consider a biometric recognition application where a video stream needs to be analyzed and a batch mode active learning algorithm has to be applied to select a batch of images to update the underlying classification model. Taking into account the specific challenges of face-based biometric data, an intuitive strategy for batch selection is to ensure that different kinds of facial appearances in a video stream are well-represented in a selected batch. This condition can be satisfied by a term which asserts that the uncertainty (entropy) of the classifier in classifying the remaining images in the video stream is minimized.

From a data geometry point of view, it is possible that the above term will only ensure that images from high-density regions are selected in the batch. This is because the set of images that are not selected may be dominated by images from such high-density regions constituting a large portion of the data. To address this issue, we introduce a term which selects images specifically from low-density regions in the data space, i.e. images that have a high distance from the remaining set.

Formally, let us consider a BMAL problem which has a current labeled set L_t and a current classifier w^t trained on L_t. The classifier is exposed to an unlabeled video U_t at time t. The objective is to select a batch B from the unlabeled stream in such a way that the classifier w^{t+1}, at time $t+1$, trained on $L_t \cup B$ has maximum generalization capability. Let C denote the possible number of classes. Then, the entropy S of the conditional distribution $P(y \mid x_j, w^{t+1})$, where x_j is the j^{th} image in the unlabeled video and y is a class label, is calculated as

$$S(y|x_j, w^{t+1}) = -\sum_{y \in C} P(y|x_j, w^{t+1}) \log P(y|x_j, w^{t+1})$$

Also, let ρ_j denote the average Euclidean distance of an unlabeled image x_j from other images in the video U_t. Greater values of ρ_j denote that the point is located in a low-density region.

The conditions described previously can thus be satisfied by defining a performance score function $f(B)$ in the following manner:

$$f(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) \tag{1}$$

The first term denotes the sum of the average distance of each selected point from other points in the unlabeled video while the second term quantifies the sum of the entropies of each remaining point in the unlabeled stream. λ_1 is a tradeoff parameter.

The problem therefore reduces to selecting a batch B of unlabeled images which produces the maximum score f(B). Let the batch size (number of images to be selected for annotation) be denoted by m, which is an unknown. Since there is no restriction on the batch size m, the obvious solution to this problem is to select *all* the images in the unlabeled video. In that case, no image will be left behind in the unlabeled video, the entropy term will become 0 and the density term will be equal to the sum of the average distances of every image from all other images and consequently, f(B) will attain its maximum score. However, querying all the images for their class labels is not an elegant solution and defeats the basic purpose of active learning. To prevent this, we modify the score function by enforcing a penalty on the batch size as follows:

$$\tilde{f}(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) - \lambda_2 m \tag{2}$$

The third term essentially reflects the cost associated with labeling the images, as the value of the objective function decreases with every single image that needs to be labeled. The extent of labeling penalty can be controlled through the weighting parameter λ_2. Defining the score function in this way ensures that any and every image is not queried for its class label. Only images for which the density and entropy terms outweigh the labeling cost term get selected.

We therefore need to select a batch B of unlabeled images so as to maximize $\tilde{f}(B)$. Since the search space is exponentially large, exhaustive search methods are not feasible. The batch selection task is therefore solved using numerical optimization techniques. Let $|U_t|$ be the number of images in the unlabeled video stream. We define a binary vector M of size $|U_t|$ where each entry M_i denotes whether the unlabeled image x_i will be selected in the current batch or not. Thus, if we want to convey the fact that image j in the video will be selected in the current batch, then M_j will be 1, otherwise M_j will be 0. We rewrite the objective function in Equation 2 into an equivalent function in terms of the defined vector M:

$$\max_{M,m} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 m \tag{3}$$

subject to the constraint:

$$M_j \in [0, 1] \tag{4}$$

To simplify the above objective function defined in terms of two variables (the vector M and the scalar m), we try to express m in terms of the other variable M. By the formulation, if an entry of M is 1, the corresponding image will be selected for annotation and if it is 0, the corresponding image will not be selected. The number of images to be selected, is therefore equal to the number of non-zero entries in the vector M, or the zero-norm of the vector M. Hence,

$$m = ||M||_0 \approx ||M||_1 = \sum_j M_j \tag{5}$$

Here, we have replaced the zero norm of M by its closest convex approximation, which is the one-norm of M. Also, from constraint 4, the one norm is simply the sum of the elements of the vector M. Substituting m in terms of M, the new optimization problem becomes:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j$$

subject to the constraint: $M_j \in [0, 1]$

The above optimization is an integer programming problem and is NP hard. We therefore relax the constraint to make it a continuous optimization problem:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{6}$$

subject to the constraint: $0 \leq M_j \leq 1$

2.2 Solving the Optimization Problem

We first write the objective function $f(M)$ as:

$$f(M) = \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{7}$$

To solve the optimization problem, we use the Quasi Newton method, which assumes that the function can be well-approximated as a quadratic in the neighborhood of the optimum point and iteratively updates the variable M to guide the functional value towards this local optima. The first derivative of the function and the Hessian matrix of second derivatives need to be computed as parts of the solution procedure. Assuming w^{t+1} remains constant with small iterative updates of M, the first order derivative vector is obtained by taking the partial of the objective with respect to M:

$$\nabla f(M_j) = \rho_j + \lambda_1 S(y|x_j, w^{t+1}) - \lambda_2$$

The Hessian starts as an identity matrix and is updated according to the BFGS method. In each iteration, a quadratic programming problem is solved which yields an update direction for M. The step size is obtained using a backtrack line search method based on the Armijo Goldstein equation [13]. The iterations are continued until the change in the value of the objective function is negligible. The final value of M is used to govern the number of points and the specific points to be selected for the given data stream (by greedily setting the top m entries in M as 1 to recover the integer solution, where $m = \sum_j M_j$). Hence, solving a single optimization problem helps in dynamically deciding the batch size as well as selecting the specific points for manual annotation.

It is to be noted that the objective function is defined in terms of the future classifier w^{t+1}, which is unknown. In the Quasi Newton iterations, w^{t+1} is approximated as the classifier trained on the current training set L_t together with the set of unlabeled points selected in the current iteration, where the label of each selected unlabeled point is assumed to be the same as that of the closest training point in L_t.

2.3 Extensions of the framework

It is straightforward to extend this formulation for dynamic batch selection to situations where multiple sources of information are available. For this, the objective function can be modified by appending relevant terms from the respective sources, together with a penalty on the batch size:

$$\begin{aligned} f(M) = \sum_{j \in U_{t1}} \rho_j M_j + \lambda_1 \sum_{j \in U_{t2}} \rho_j M_j - \lambda_2 \sum_{j \in U_{t1}} (1 - M_j) S(y|x_j, w^{t+1}) \\ -\lambda_3 \sum_{j \in U_{t2}} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_4 \sum_j M_j \end{aligned}$$

Moreover, if contextual information is available (eg location of a subject, whether at home or in office), it can be used to construct a prior probability vector depicting the chances of seeing particular acquaintances in a given context. The entropy term can then be computed on the posterior probabilities obtained by multiplying the likelihood values returned by the classifier with the context aware prior. Thus, subjects not expected in a given context (eg. a home acquaintance in an office setting) will have low priors and consequently, the corresponding posteriors will not contribute much in the entropy calculation. The framework can therefore be extended to perform context-aware adaptive batch selection.

3 Experiments and Results

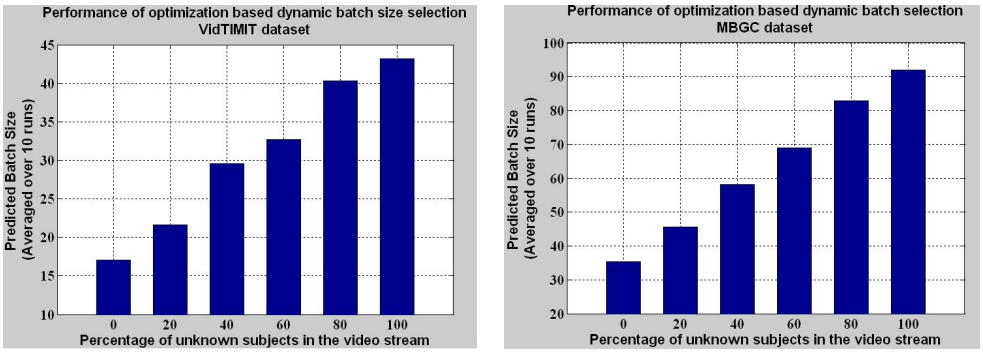
3.1 Datasets and Feature Extraction

To demonstrate the effectiveness of the framework on biometric video data, we used the VidTIMIT and the MBGC (Multiple Biometric Grand Challenge) datasets in our work. The VidTIMIT dataset contains video recordings of subjects under natural conditions. The MBGC is the leading dataset for biometric recognition collected by the National Institute of Standards and Technology (NIST) and contains video recordings of subjects under uncontrolled indoor and outdoor lighting. 25 subjects were randomly chosen from each dataset for our experiments. Our preliminary experiments (not presented here due to lack of space) confirmed that the Discrete Cosine Transform (DCT) feature could effectively distinguish the different subjects and hence this was used in our work.We carried out two experiments to demonstrate the effectiveness of our framework - the first was performed

to study the usefulness of dynamic batch size selection and the second was conducted to show the superiority of the proposed approach in selecting unlabeled data points as compared to heuristic BMAL techniques.

3.2 Experiment 1

The 25 subjects in each dataset were randomly divided into 2 groups - a "known" group containing 20 subjects and an "unknown" group containing the remaining 5 subjects. The learner was trained with 1 video of each of the "known" subjects. Unlabeled video streams were then presented to the active learner and it was asked to query images based on the proposed framework. The percentage of unknown subjects in the unlabeled video stream was gradually increased from 0% to 100% in steps of 20%. However, the learner was not given any information about the composition of each unlabeled stream. Each unlabeled video had 100 images in total to facilitate fair comparison. The



(a) Dynamic Batch Size Selection on the VidTIMIT dataset

(b) Dynamic Batch Size Selection on the MBGC dataset

Figure 1: Optimization based dynamic batch size selection on the VidTIMIT and MBGC datasets.

results on the VidTIMIT and the MBGC datasets are shown in Figure 1 (λ_1 and λ_2 were both empirically set to 1 in this experiment). The x axis depicts the percentage of unknown subjects in the video stream and the y axis represents the batch size as decided by the active learner. Also, each bar represents the average performance over 10 trials with different images of the known and unknown subjects to rule out the effects of randomness. It is noted that, in both the datasets, as the proportions of unknown subjects in the supplied video increases, the learner decides on a larger batch size. This matches our intuition because, with growing proportion of unknown subjects, the confidence of the learner on the video stream decreases and so it needs to query a larger number of images to achieve good generalization capability. Hence, the framework enables the active learner to automatically and adaptively choose the batch size based on its level of uncertainty on the images of a given video stream. This corroborates the effectiveness of the optimization framework in dynamic batch selection. Similar results were obtained (not reproduced here for the sake of brevity) when the unlabeled video contained images of subjects with varying poses and expressions as compared to the training set.

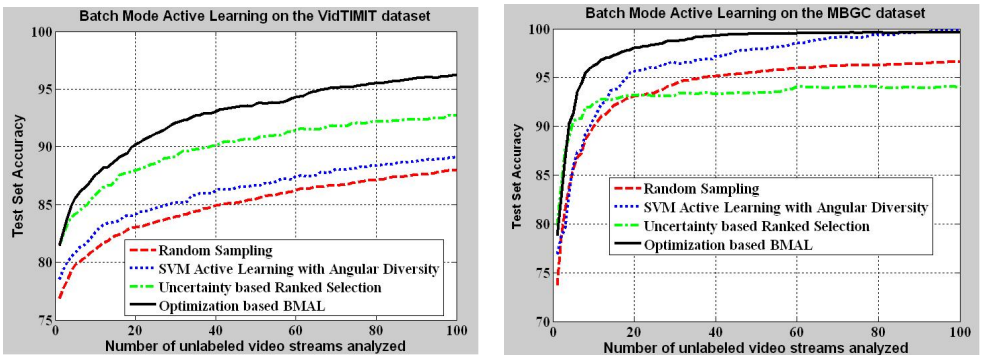
3.3 Experiment 2

In this experiment, a classifier was induced with 1 training video of each of the 25 subjects. 100 unlabeled video streams were then presented to the classifier one after another. The images in the video streams were randomly chosen from all 25 subjects and did not have any particular proportion of subjects in them. This was done to emphasize the performance of the framework under general conditions. For each stream, the batch size was dynamically selected and optimization based BMAL was used to select a batch of images. The selected images were appended to the training set, the classifier updated and then tested on a test video containing 4500 images spanning all the 25 subjects. The objective was to study the growth in accuracy on the same test video with increasing size of the training set.

The proposed optimization based approach was compared with three other BMAL schemes - (i) Random Sampling (ii) SVM Active Learning with Angular Diversity, where a batch of points was incrementally sampled such that at each step the hyperplane induced by the selected point maximizes the angle with all the hyperplanes of the already selected points, as proposed by Brinker [7] and

(iii) Uncertainty Based Ranked Selection, where the top k uncertain points were queried from the unlabeled video, k being the batch size.

For each video stream, the dynamically computed batch size was noted and used for the corresponding unlabeled video in each of the heuristic techniques, for fair comparison. The results are shown in Figure 2. As the x axis of the graphs indicate, with every new unlabeled video stream that enters the system, the performance of the classifier improves over time. It is noted that the proposed optimization based framework performs much better than the other methods as its accuracy on the test set grows at the fastest rate. The label complexity (the number of labeled examples needed to achieve a certain accuracy) is least in case of the proposed technique. This corroborates the superiority of the framework over other similar techniques, under general conditions which reflect real world scenarios.



(a) BMAL on the VidTIMIT dataset (b) BMAL on the MBGC dataset

Figure 2: Performance of different BMAL strategies on the VidTIMIT and MBGC datasets.

4 Conclusion

In this paper, we introduced a novel optimization based framework to dynamically select the batch size in batch mode active learning applications. Using a penalty term on the batch size in the objective function, our framework can simultaneously solve for the batch size as well as the specific points to be selected. The computational complexity of our method is the same as the best performing static BMAL algorithm [12] where the batch size needs to be specified in advance. The results certify the potential of this approach in being used in any BMAL problem.

References

[1] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," in *JMLR*, 2000.

[2] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *JAIR*, 1996.

[3] Y. Freund, S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning.*, 1997.

[4] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," *ICAI*, 1997.

[5] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *JMLR*, vol. 5, 2004.

[6] A. McCallum and K. Nigam, "Employing EM and Pool-Based active learning for text classification," in *ICML*, 1998.

[7] K. Brinker, "Incorporating diversity in active learning with support vector machines," *ICML*, 2003.

[8] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *International Conference on World Wide Web*. ACM, 2006.

[9] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *IEEE CVPR*, 2008.

[10] S. Hoi, R. Jin, and M. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE TKDE*, 2009.

[11] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *ICML*, 2006.

[12] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *NIPS*, 2008.

[13] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.