Cost-Sensitive Batch Mode Active Learning: Designing Astronomical Observation by Optimizing Telescope Time and Telescope Choice

Xide Xia* Pavlos Protopapas† Finale Doshi-Velez‡

Abstract

Astronomers and telescope operators must make decisions about what to observe given limited telescope time. To optimize this decision-making process, we present a batch, cost-sensitive, active learning approach that exploits structure in the unlabeled dataset, accounts for label uncertainty, and minimizes annotation costs. We first cluster the unlabeled instances in feature space. We next introduce an uncertainty-reducing selection criterion that encourages the batch of selected instances to span multiple clusters, in addition to taking into account annotation cost. Finally, we extend this criterion to incorporate the fact that nearby astronomical objects may be observed at the same time. On two large astronomical data sets, our approach balances the trade-offs among FOV, aperture, and time cost and, therefore, helps astronomers design effective experiments.

1 Introduction

Limited observing time is a bottleneck for many astronomy studies. Important questions in study design include what telescope to use and how to best take advantage of limited observing time. Currently, a Time Allocation Committee (TAC) allocates observing time based on the scientific justification for the proposal and the observational planning required to ensure a smooth running of the telescope. The odds of proposal acceptance depends additionally on the demand, and thus it is harder to obtain time on more powerful telescopes (large aperture or large field of view (FOV)). In this work, we argue that choosing the next observation based on the results of the previous observation, rather than having a fixed plan at the beginning of the experiment, may assist in utilizing telescopes efficiently.

Active learning is a classification strategy that iteratively selects a subset of data for annotation from a large amount of unlabeled data. A classifier trained on that subset of data is used to label the remaining subset. To date, most research in active learning has focused on iteratively selecting the next single instance to label [5, 16, 17, 7, 29, 26, 9, 15]. However, in the astromony setting, one must make decisions about how to use an entire block of observing time in advance, as it may not be possible to analyze the data and change the telescope pointing in real-time. The data collected in that batch can be used to inform the observation choices for the next observation block. This process is called batch-mode active learning [14]; it is more challenging than the single-instance case because we now must address potential information overlap between selected instances.

While observing time for specific experiments is scarce, the advent of CCDs and modern computers has enabled several large-scale sky surveys; (MACHO) [1, 28] and (EROS) [21] have observed tens of millions of stars and other objects in the direction of the Large and Small Magellanic Clouds, producing hundreds of terabytes of light curves. The Sloan Digital Sky Survey [12] and PanSTARRS [13] have repeatedly observed large parts of sky for many years, and the Large Synoptic Survey Telescope [27] (the largest and most ambitious project ever in astronomy) a survey under construction at the Atacama desert in Chile, will be observing close to a billion objects every few nights for many years. The data from these surveyes—which are designed to answer as many astronomical questions as possible—are often not suitable for the more specific questions posed by any particular study. For example, object classification requires manual inspection of observations at several wavelengths, not all of which may have been collected in the survey.

In this paper, we present an active learning approach that allows us to leverage these vast sources of unlabeled astronomy data. Specifically, we introduce *Cost-Sentitive Batch Mode Active Learning* (**C-BAL**), a batch mode active learning approach that minimizes both observation costs and label uncertainty by taking advantage of the geometry of the unlabeled data set. Specifically, our approach combines the following char-

*Institute For Applied Computational Science, John A. Paulson School of Engineering and Applied Science, Harvard University, xidexia@g.harvard.edu.

†Institute For Applied Computational Science, John A. Paulson School of Engineering and Applied Science, Harvard University, pavlos@seas.harvard.edu.

‡John A. Paulson School of Engineering and Applied Science, Harvard University, finale@seas.harvard.edu.

acteristics:

- Minimizes Cost. We consider the observation cost (telescope time) associated with gathering the data. Our approach also leverages the fact that multiple objects may be present in the same frame, or *pointing*, and thus can be analyzed and labeled for a single observation time cost.
- Leverages Representation. When selecting objects, we take into account the geometry of the data set by clustering the unlabeled objects and choosing objects from different clusters. Using an unsupervised method to guide the active learning helps us select objects that may potentially assist in labeling many uncertain points (e.g. from a large cluster) and avoid selecting multiple similar uncertain objects in one batch. A clustering-based approach to grouping uncertain points is also sensible given our choice of random forest classifier, which does not have a global form for its decision-boundary; however, one can reasonably assume that locally similar objects would be classified with the same label.
- Minimizes Expected Label Uncertainty. As with other active learning approaches, we select objects from the set of objects with high label-uncertainty. There are many ways to measure the uncertainty. In this paper, we choose random forest classifier so that the uncertainty of instances can be simply measured by the empirical proportion of decision trees that make the same prediction.

We apply our approach to classifying objects from the MACHO and EROS databases. In this application, getting labels for objects first requires pointing the telescope to that section of the sky to collect measurements. An astronomer must submit a set of observations that he or she wishes to make in a particular time window; once these are complete the astronomer may analyze the results and request additional observations. This active learning problem has several important characteristics: First, different objects may have the same label-uncertainty, but a brighter object will require less time to label than a dimmer object. Second, each telescope can only see a part of the sky for a single observation. Since all objects within that FOV are captured, not just the object of interest, several objects may be labeled at once.

Compared to prior work in active learning for astronomy [31, 24], our approach balances trade-offs between expected information reduction and time cost by not only exploiting the geometry of the data set, but also taking the opportunity to gather several labels at once into consideration. (That is, it takes advantage of a telescope's FOV to increase the total expected uncertainty reduction of each batch of selected objects.) By balancing the size of FOV, aperture, and annotation time cost budget, we expect our approach will assist astronomers in designing effective experiments.

2 The Active Learning Framework

Single-instance Active Learning Starting with a set of unlabeled instances or a few labeled instances, the goal of active learning is to sequentially request as few labels as possible to achieve high classification performance. Suppose we have a labeled set L of K input features x_k and labels y_k: $\{(x_1, y_1), \ldots, (x_K, y_K)\}$. Let the set U consist of the remaining unlabeled data. Our goal is to select the next instance x_{K+1} to label to minimize anq expected loss $\ell(\cdot)$ on the remaining data $x_n \in U$:

$$\min_k \quad \mathbb{E}_{y_K, y_n \in U}[\sum_{x_n \in U} \ell(\hat{y_n}, y_n)] \tag{2.1}$$

where $\hat{y_n} = f_s(x_n)$, $f_S(x)$ is the classifier trained with the labeled set L.

Equation 2.1 is generally intractable to optimize directly. Common methods to choose the next point x_{K+1} include querying the instance that maximally reduces the uncertainty in the labels $\hat{y}_n$ [5], [16], [17], that is closest to the current decision boundary [7], [29], [26], and that has the most disagreement among different classifiers in an ensemble [9], [15]

Batch-mode Active Learning Batch mode active learning selects a group of instances to label at each iteration. Again, heuristics must be used to select instances as the optimization of selecting a group to add is even more challenging than the single instance case; in particular, we must not choose instances that would give us essentially the same information. While joint uncertainty reduction heuristics are used [32], a popular approach is to choose the group $x_{K+1}, x_{K+2}, \cdots$ in a way such that it somehow covers the unlabeled data [6, 11] or such that the distribution over the features x of the labeled and unlabeled data remains similar [8].

Budgeted Active Learning In budgeted or cost-sensitive active learning, each instance x_n has a labeling cost $c(x_n)$. Suppose that we are given some budget B; our goal is to iteratively select instances into the labeled set L to

$$\begin{aligned} &\min_k && \mathbb{E}_{y_K, y_n \in U}[\sum_{x_n \in U} \ell(\hat{y_n}, y_n)] \\ &\text{subject to} && c(L) \leq B. \end{aligned} \tag{2.2}$$

3 A Framework of Batch Mode Cost-Sensitive Active Learning.

We focus on the case of cost-sensitive batch model active learning, which occurs when astronomers must select instances to classify astronomical objects with limited observation time. Specifically, we assume that in every iteration, we are given a budget B—some amount of observing time—that we must allocate to specific instances, and our goal is to optimize Equation 2.2. Because Equation 2.2 is intractable to optimize directly, we propose a surrogate score function that encourages the selection of different instances. We show that this problem has the form of the submodular knapsack problem, which allows us to efficiently select a good batch within our budget B.

3.1 Score Function for a single x: Expected Uncertainty Reduction We start with the simpler problem of choosing one instance at a time, when all instances are of equal cost. Our goal is to choose instances to minimize the total label uncertainty across all the unlabeled instances

$$\min_{x_k} \mathbb{E}_{y_k}[\sum_{x_n \in U \cup x_k} 1 - \max(p(\hat{y}_n))] \tag{3.3}$$

where $p(\hat{y}_n)$ is the probability of the most probable label for the observation x_n.

Solving Equation 3.3 is intractable; to approximate it we first cluster the unlabled observations $x_n \in U$. Next, we posit that each point x_k in the cluster c, if it were to be labeled, will label some proportion r of its cluster, depending on on how close it is to the center of its cluster:

$$r(x_k, c_{x_k}) \propto \exp(\text{-dist}(x_k, c_{x_k})) \tag{3.4}$$

where c_{x_k} is the center of the cluster of x_k.

Thus, the total expected uncertainty reduction, if we label an instance x_k, is given by

$$\text{score}(x_k) = r(x_k, c_{x_k}) \sum_{x_i \in c} 1 - \max(p(\hat{y}_i)) \tag{3.5}$$

This score function takes into the account the probability that a cluster will be labeled, the size of the cluster, and the uncertainty of the unlabeled elements in the cluster.

3.2 Score Function for a Pointing (Multiple Simultaneous Labels) In the astronomy domain, we may have the opportunity to collect multiple labels at the same time: all astronomical objects in the same FOV of the telescope can be observed simultaneously. Astronomers refer to the process of pointing the telescope towards a target and taking an image as *pointing*. Thus, the problem of selecting an instance x_k becomes selecting a pointing p_i. The expected uncertainty reduction of a pointing can be given by

$$\text{score}(p_i) = \sum_{c \in p_i} \max_{x_k \in c \cup p_i} (\text{rep}(x_k, c)) \sum_{x_i \in c} 1 - \max(p(\hat{y}_i)) \tag{3.6}$$

where $c \in p_i$ are all the clusters associated with the observations in the pointing p_i.

Algorithm 1 Batch Mode Pointing Selection based on C-BAL

Input: The labeled instances $\boldsymbol{L}$, unlabeled instances $\boldsymbol{U}$, the size of telescope (FOV) d, and the budget $\boldsymbol{B}$.
Output: A batch of selected pointings $\boldsymbol{P}$.

1: Compute the uncertainty for each instance.
2: Implement K-Means clustering with $K = b$ on $\boldsymbol{U}$.
3: **for** x_i in $\boldsymbol{U}$ **do**
4: Set x_i as the center of a new created pointing $\boldsymbol{p}_i$. Find all the instances $\boldsymbol{X}_i$ inside a FOV of d degree. Add x_i to $\boldsymbol{X}_i$.
5: Compute the score of $\boldsymbol{p}_i$ as Equation 3.6.
6: **end for**
7: Let $\boldsymbol{P}_0 = p_0$ and $\boldsymbol{I}_0 = \boldsymbol{X} \setminus p_0$, where p_0 is the poingting with highest score. c_0 is the observation cost of p_0 .
8: **while** $\boldsymbol{I}_{t-1} \neq \emptyset$ **and** $\sum_{i \in \boldsymbol{P}_{t-1}} c_i < \boldsymbol{B}$ **do**
9: Find $\arg\max_i \frac{score(L \cup \boldsymbol{P}_{t-1} \cup p_i) - score(L \cup \boldsymbol{P}_{t-1})}{c_i}$.
10: Let $\boldsymbol{P}_t = \boldsymbol{P}_{t-1} \cup p_i$ and $\boldsymbol{I}_t = \boldsymbol{I}_{t-1} \setminus p_i$ if $\sum_{i \in \boldsymbol{P}_{t-1} \cup p_i} c_i \leq \boldsymbol{B}$. Otherwise, let $\boldsymbol{P}_t = \boldsymbol{P}_{t-1}$ and $\boldsymbol{I}_t = \boldsymbol{I}_{t-1} \setminus p_i$.
11: **end while**
12: Return $\boldsymbol{P}_t$

3.3 Score function for a batch of pointings When we are choosing a batch of pointings $\{p_i\}$, the problem is quite similar to the single pointing selection. The expected uncertainty reduction of $\{p_i\}$ can be given by

$$\text{score}(\{p_i\}) = \sum_{c \in \{p_i\}} \max_{x_k \in c \cup \{p_i\}} (\text{rep}(x_k, c)) \sum_{x_i \in c} 1 - \max(p(\hat{y}_i)) \tag{3.7}$$

where $c \in \{p_i\}$ are all the clusters associated with the observations in the bath of pointings $\{p_i\}$.

3.4 Optimizing the Score In the simplest case, the costs of all pointings $c(p_i)$ are identical. We first show

that Equation 3.7 is submodular:

$$\text{score}(\{x_i\}) + \text{score}(\{x_j\}) \geq \text{score}(\{x_i\} \cup \{x_j\}) + \text{score}(\{x_i\} \cap \{x_j\}) \tag{3.8}$$

where $\{x_i\}$ and $\{x_j\}$ two different set of instances. Thus, to select a near-optimal batch with budget B, we can simply greedily select pointings p_i until our budget is exhausted.

However, in astronomy, the costs $c(p_i)$ are not identical—dimmer objects require more observation time to label than brighter objects. We assume that we are given a budget B of observing time, and we are tasked to use it to maximize the score. This corresponds to the budgeted submodular knapsack problem, and can be solved near-optimally by starting with all subsets of single and pairs of pointings and then greedily adding pointings according to

$$\arg\max_i \frac{score(L \cup p_i) - score(L)}{c(p_i)} \tag{3.9}$$

until the observing budget B is exhausted [25], see Algorithm 1.

Table 1: MACHO Data Set Composition

	Class	Number of objects
1	Non variable	966
2	Quasars	59
3	Be Stars	101
4	Cepheid	610
5	RR Lyrae	255
6	Eclipsing Binaries	126
7	MicroLensing	580
8	Long Period Variable	365

Table 2: EROS Data Set Composition

	Class	Number of objects
1	BV	829
2	CEP	1500
3	DSCT	1114
4	EB	1484
5	LPV	1500
6	QSO	251
7	RRLYR	1499
8	T2CEP	123

4 Experiment Setup

4.1 Data Sets We conduct experiments on two astronomical survey databases, MACHO and EROS. MACHO (Massive Compact Halo Object)[1] observed the sky from 1992 to 1999 to detect microlensing events produced by the Milky Way halo objects. Several tens of millions of stars were observed in the Large Magellanic Cloud, Small Magellanic Cloud, and Galactic bulge. We use a subset of 3063 labeled observations from the MACHO catalog [19, 20, 18, 31], which consists of several sources from MACHO variable studies [1]. EROS (Exprience pour la Recherche d'Objets Sombres) is another survey whose objective is the search and study of dark stellar bodies that belong gravitationally to our galaxy. We use a subset of 8317 labeled observations [2]. The class proportions of the MACHO and EROS datasets are presented in Tables 1 and 2 respectively. For both of MACHO and EROS, every light curve is described by a vector of 64 features generated using the *Feature Analysis for Time Series* library (FATS) [23].

4.2 Evaluations

4.2.1 Select Single Astronomical Objects In astronomy, brightness of an astronomical object is measured by a quantity known as magnitude (*mag*). Magnitude is a negative logarithmic measurement of the brightness of a star (subtracted from a arbitrary constant). The larger the magnitude, the dimmer the star is. The observation time required to achieve some predefined signal-to-noise ratio for an object is proportional to its magnitude:

$$\boldsymbol{C}(x_i) \propto 10^{0.4 mag_{x_i}}. \tag{4.10}$$

The labeling cost of a batch of instances $\{x_i\}$ is the sum of the time cost of each selected instances:

$$\boldsymbol{C}(\{x_i\}) \propto \sum_{x_i \in \{x_i\}} 10^{0.4 mag_{x_i}}. \tag{4.11}$$

We optimize equation 4.11 with 1.

To test the performance of proposed method, we randomly split each data sets into two parts: a train set from where models select instances and a test set which is used to test the performance of models. For MACHO, the sizes of two sets were 2000 and 1063. For EROS, the sizes of two sets were 6000 and 2317. For each data set, we randomly select 20 instances from the unlabeled pool as initial labeled data. At each iteration of active learning, 10 instances are selected by the active learning methods, and then added into the labeled data. The reported results were compiled as the average of experiments run 10 times.

We tested our approach with and without considering the annotation cost (**C-BAL Without Cost** and **C-BAL**). We compared the performance of the proposed algorithm with the following active learning approaches:

Rand: the baseline that selects instances randomly.

Top-Uncertainty: uncertainty sampling that decides the cases whose labels are the most uncertain for the current classifier.

S1: method proposed in [24], which selects instances whose feature density is the most under-sampled by the training data.

S2: method proposed in [24], which selects instances that maximize the total amount of change in the predicted probabilities.

Zhu: method proposed in [34], which combines active learning and semi-supervised learning using gaussian fields and harmonic functions.

C-BAL Without Cost: our approach without considering the observing cost, Equation 3.5.

C-BAL: our approach taking the observing cost into consideration.

During clustering, we determine the number of clusters by maximizing the Bayesian information criterion (BIC) [22]:

$$BIC = 2\log p(x|\boldsymbol{M}) + k\log(n) \tag{4.12}$$

where $\log p(x|\boldsymbol{M})$ is the maximized value of the likelihood of the model $\boldsymbol{M}$, k is the number of free parameters to be estimated, and n is the number of data points. Using the BIC criterion, we set the number of clusters on MACHO and EROS are 100 and 120 respectively. The number of clusters was held constant in each iteration.

4.2.2 Selecting Pointings When observing a star, we can get information about the other objects in the same pointing. We consider two scenarios: in the first, when we choose a pointing, we assume that we will collect data on all the objects in the pointing excluding outliers (defined as the dimmest 15% of objects). In the second, we optimize over both pointings and observation times; that is, we may choose not to observe all of the objects in a pointing.

4.2.3 Selecting Telescopes So far, we have discussed how FOV can affect the performance of learning by observing multiple astronomical objects at the same time. Now, we include the *aperture*, another important factor that can influence our observing time. Aperture is the size of the telescope's primary mirror, which determines how much light reaches the image plane. For the same brightness, a wider aperture requires less time than a narrow one. We write the time required to observe a star x with a telescope T as

$$\boldsymbol{C}(\boldsymbol{x}, \boldsymbol{T}) = \frac{\mu}{R_T^2} 10^{0.4 mag_x}. \tag{4.13}$$

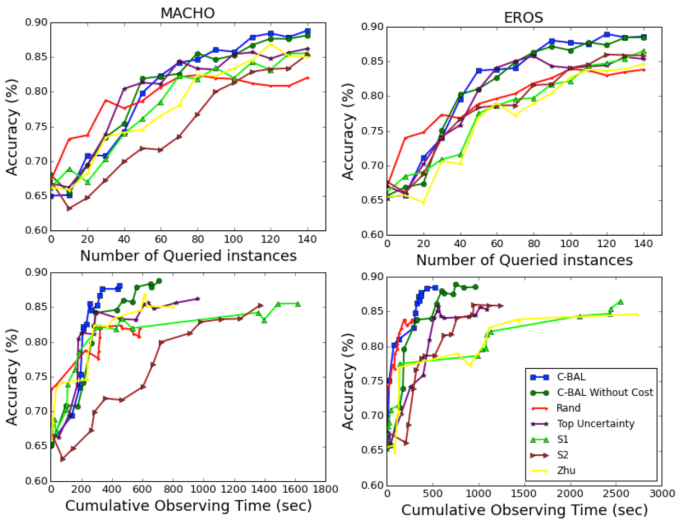

Figure 1: Comparative performance of different active learning methods on MACHO and EROS data sets.

where R_T is the aperture (diameter) of T, mag_x is the magnitude of x, and μ is a constant value that depends on the sensitivity of the sensor, reflectivity of the mirrors and other factors. For simplicity, we use $\mu = 1.07568 * 10^{-7}$ [sec $\cdot$m^2] for all telescopes derived from Canada France Hawaii Telescope with megacam [1].

We collect information of different telescopes and for each telescope, we use their real aperture and FOV. Using these real dataset, we conduct experiments of the proposed method on MACHO data to compare the performance of different telescopes.

5 Results

5.1 Time Cost of Observing Single Astronomical Object The results of single object selection by using different active learning approaches on the MACHO and EROS data sets respectively are showed in Figure 1. In the upper panel, we show the accuracy as a function of the number of queried instances; and in the lower panel, we evaluate the performance as a function of the cumulative observation time cost. When only considering the total number of queried instances, the C-BAL Without Cost outperforms C-BAL because there is no penalty for choosing an expensive instance; C-BAL Without Cost chooses the instances which have the highest total expected uncertainty reduction. However. C-BAL is more effective when observation time budget is limited (lower panel).

[1] http://www.cfht.hawaii.edu/

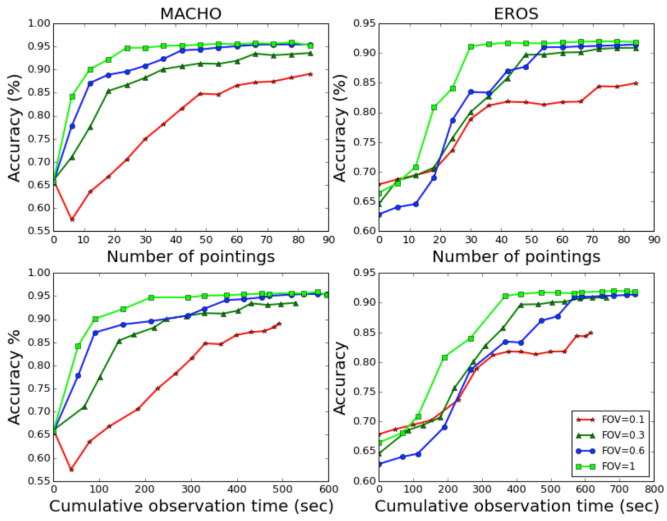

Figure 2: Comparative performance of different FOVs from 0.1 to 1 degree on MACHO and EROS data sets.

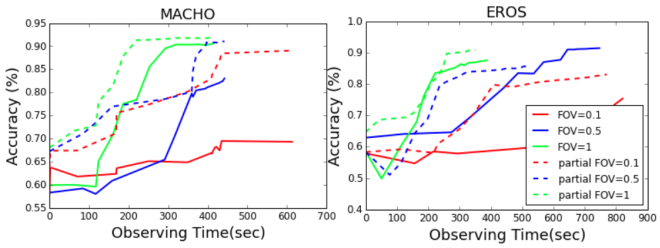

Figure 3: Benefits of allowing for partial pointings. Left: Result on the MACHO data set. Right: Result on the EROS data set.

5.2 Time Cost of Observing Pointings We compare the performance of C-BAL under different telescope FOVs. Here, we set aperture to be a fixed value (default =1). Figure 2 shows the results of comparing the performance of C-BAL under different FOVs from 0.1 to 1 degree[2]. For each FOV, we have run the same optimization for 15 times and presented the average value. The results show that, when a fixed time budget is given, a larger FOV tends to have better performance while a smaller FOV results in a slower accuracy improvement. A contour plot of the accuracy as a function of time cost and FOV on the MACHO and EROS data sets is included in the supplementary materials.

Also, instead of observing all objects inside one pointing $\boldsymbol{p}$, we are able to observe part of them so that we can save telescope time. For example, suppose that observing all objects inside one pointing p will take 5 seconds. However, we might be able to ignore the 40% dimmest ones and observe the 60% left inside p within 4 seconds. In such cases, we can save telescope time by losing information from those dimmest objects. As before, we solve for the the trade-off between saving time cost and partly losing expected uncertainty reduction can be solved by the approximation to the submodular knapsack problem, except now we must solve over (pointing, observing time) pairs. The results shown in Figure 3 demonstrate that allowing for partial observations performs significantly better than always trying to resolve all of the objects in a pointing.

[2]FOV units are in degrees

5.3 Application to Choosing Real-World Telescopes To test performance of C-BAL on optimizing real-world telescope choice, we collect the information of 11 telescopes all over the world (see Table 3). Besides FOV, the aperture information of telescopes is also shown. In optics, an aperture is an opening through which light travels and it is one of the important parameters that determine the observation time cost of the telescope. The time cost is formulated as Equation 4.13.

We show the performance of C-BAL on each telescope in Table 3 on MACHO data set. For each telescope, we select 6 pointings at each iteration using the method described in Algorithm 1. The observation time cost is determined by the corresponding aperture of the telescope and the highest magnitude value of all instances inside the FOV. Figure 4 shows the contour plot of accuracy as a function of FOV and aperture when different time cost budgets are given: 50, 100, and 500 seconds respectively. Besides, we also pin the performance of telescopes in each plot and the annotation labels are the corresponding index of telescopes in Table 3. In the figure, a telescope with a higher degree of FOV or larger aperture has better performance when the observation time budget is low. However, the accuracy improvement tends to slow down when the observation time budget is large enough. Figure 4 helps to find the best choice of telescope when the time budge is given. For example, when the time budget is low (say 50 seconds), we will have to choose wide field telescopes such as Blanco 4m-DECam telescope in Chile to achieve a high accuracy. However, if we have a larger time budget, we will have much more choices.

Run time. In general, astronomers may have days to weeks to decide the next set of pointings, so the computational running time of the active learning algorithm is not a key factor. That said, our approach runs reasonably fast, even for these large astronomy data sets: for the MACHO data set, which has 3063 objects, our non-optimized Python implementation of our algorithm re-

Table 3: Real-World Telescope Data

	Country	Observatory	Telescope	Instrument	Apperture	FOV (arcmin)
1	Chile	Las Campanas	Bode	Mega Cam	6.1	25*25
2	Chile	Las Campanas	Bode	IMACS	6.1	21*21
3	Chile	Cerro Tololo	Blanco 4m	ISPI	4	10.25*10.25
4	Chile	Cerro Tololo	Blanco 4m	DECam	4	132*132
5	Chile	Cerro Tololo	SOAR	SOAR Imager (SOI)	4.1	5.2*5.2
6	Chile	Cerro Tololo	SMARTS 1m	Y4K CAM	1	20*20
7	Tuscon	Kitt Peak	Mayall 4m	KOSMOS	3.7	36*36
8	Hawaii	Mauna Kea	CFHT	Mega Cam	3.6	60*60
9	Chile	La Silla	MPG/ESO 2.2-metre	WFI	2.2	33*33
10	Chile	Paranal	Very large telescopeX4	MUSE	8.2	1*1
11	Chile	Gemini	Gemini 9m	GSAOI	9	1.333*1.333

quired 1,688 seconds on a standard laptop for an observation budget $B = 50$ seconds; 2,792 seconds for B=100 seconds; and 7,823 seconds for B=500 seconds.

6 Related Work

A large number of strategies for active learning have been proposed for classification in recent years. One common strategy is based on uncertainty sampling [5, 16, 17, 7, 29, 26, 9, 15], which always aims to query the instance with least certain. Another popular approach is combining active learning and semi-supervised learning [34] so that one can efficiently estimate the expected generalization error after querying a point. [34] introduced an active learning framework based on Gaussian random fields and harmonic functions. However, in the astronomy setting, this becomes unrealistic - one must make decisions about how to use an entire block of observing time in advance.

Because of these challenges, in most astronomy applications, experimental designs are decided beforehand, without active learning. Closest to our work are [31, 24]. [31] presented a Bayesian nonparametric approach to selecting filters for sequential experimental design for astronomical observations. And [24] compared several different methods to avoid the problem of sample selection bias, which could cause significant errors in predictions on the testing data. The results from [24] showed that two of the active learning selection criteria proposed have the best performance. The former criterion is to measure by density, and the latter one incorporates both diversity and uncertainty.

In this paper, we present a batch mode active learning approach that allows us to leverage vast sources of unlabeled data. We minimize both observation costs and label uncertainty by taking advantage of the geometry of the unlabeled data set.

7 Conclusion

In this paper, we proposed a batch mode cost-sensitive active learning approach, C-BAL, that incorporated label uncertainty, representativeness, and observation cost. We first introduced that observing time cost of labeling astronomical objects varied depending on the flux magnitudes of the celestial objects. We then conducted the experiments on two astronomical databases: MACHO and EROS. Our strategy out-performed several standard approaches to batch-mode active learning when considering the limitation of an observing time budget. We then extended C-BAL to situations in which we could observe more than one unlabeled instance at one time given the Field of View (FOV) of a telescope. The results balanced the trade-off between the FOV and observing time cost budget. Finally, we applied our proposed approach to 11 real-world telescopes and evaluated their performance on the MACHO data. This practical application may help astronomers optimize their telescope choice when observing time budget is limited as well as estimate observing time cost when the telescope is given.

References

[1] C. Alcock, R.A. Allsman, D. Alves, T.S. Axelrod, D.P. Bennett, K.H. Cook, K.C. Freeman, K. Griest, J. Guern, M.J. Lehner, S.L. Marshall, H.-S. Park, S. Perlmutter, B.A. Peterson, M.R. Pratt, P.J. Quinn, A.W. Rodgers, C.W. Stubbs, and W. Sutherland. "The MACHO project: 45 candidate microlensing events from the first year galactic bulge data." In *The Astrophysical Journal*, 479(1): 119, 1997.

[2] Alcock, Ch, et al. "EROS and MACHO combined limits on planetary-mass dark matter in the galactic halo." The Astrophysical Journal Letters 499.1 (1998): L9.

[3] C. Alcock, R.A. Allsman, D. Alves., et al. "The MACHO project: limits on planetary mass dark matter in

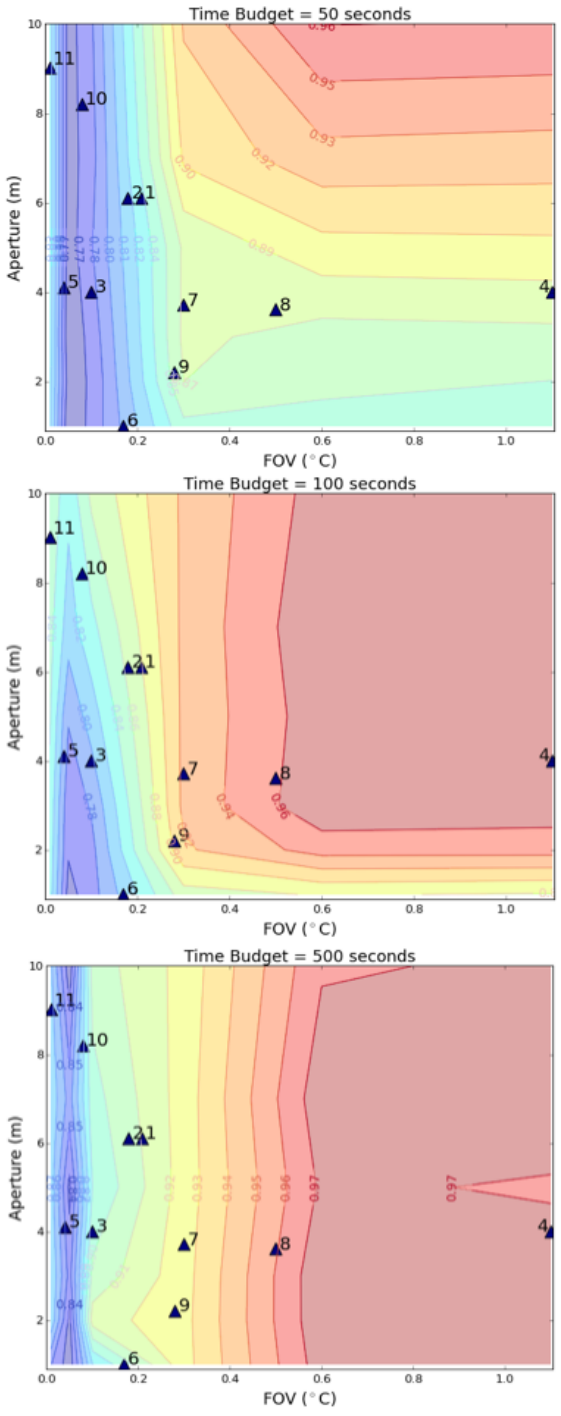

Figure 4: Contour plot of accuracy as a function of FOV and aperture with different time cost budgets.

the galactic halo from gravitational microlensing." In *The Astrophysical Journal*, 471(2): 774, 1996.

[4] J. Attenberg, P. Melville, and F. Provost. "Guided feature labeling for budget-sensitive learning under extreme class imbalance." In *ICML* Workshop on Budgeted Learning. 2010.

[5] J. Attenberg and F. Provost. "Inactive learning?: difficulties employing active learning in practice." In *ACM SIGKDD Explorations Newsletter.*, 12(2): 36-41, 2011.

[6] K. Brinker. "Incorporating diversity in active learning with support vector machines". In *ICML*. 3: 59-66, 2003.

[7] C. Campbell, N. Cristianini, and A. Smola. "Query learning with large margin classifiers." In *ICML*. 111-118, 2000.

[8] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson., S. Panchanathan, and J. Ye. "Batch mode active sampling based on marginal probability distribution matching." In *ACM Transactions on Knowledge Discovery from Data (TKDD).*, 7(3): 13, 2013.

[9] I. Dagan and S.P. Engelson. "Committee-based sampling for training probabilistic classifiers." In *Proceedings of the Twelfth International Conference on Machine Learning*. 150-157, 1995.

[10] P. Donmez, J.G. Carbonell and P.N. Bennett. "Dual strategy active learning." In *ECML* 2007. Springer Berlin Heidelberg 116-127, 2007.

[11] P. Donmez and J.G. Carbonell. "Paired-Sampling in Density-Sensitive Active Learning." 2008.

[12] D.J. Eisenstein, D.H. Weinberg, E. Agol, et al. "Sdss-iii: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems." In *The Astronomical Journal* 142(3): 72, 2011.

[13] K.W. Hodapp, N. Kaiser, H. Aussel, et al. "Design of the Pan-STARRS telescopes." *Astronomische Nachrichten*, 325.68 (2004): 636-642.

[14] S.J. Huang and Z.H. Zhou. "Fast multi-instance multi-label learning." arXiv preprint arXiv:1310.2049, 2013.

[15] S.J.Huang, R. Jin and Z.H. Zhou. "Active learning by querying informative and representative examples." Advances in neural information processing systems. 892-900, 2010

[16] F. Jing, M. Li, H.J. Zhang, et al. "Entropy-based active learning with support vector machines for content-based image retrieval." In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on. IEEE* 1: 85-88, 2004

[17] A.J. Joshi, F. Porikli and N. Papanikolopoulos. "Multi-class active learning for image classification." In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE.* 2372-2379, 2009

[18] D.W. Kim, P. Protopapas, C.A. Bailer-Jones, Y. Byun, S. Chang, J. Marquette, and M. Shin. "The EPOCH Project-I. Periodic variable stars in the EROS-2 LMC database." In *Astronomy & Astrophysics*, 566: A43, 2014.

[19] D.W. Kim, P. Protopapas, Y. Byun, C. Alcock, R. Khardon, and M. Trichas. "Quasi-stellar object selec-

tion algorithm using time variability and machine learning: Selection of 1620 quasi-stellar object candidates from MACHO Large Magellanic Cloud database." In *The Astrophysical Journal*, 735(2): 68, 2011.

[20] D.W. Kim, P. Protopapas, M. Trichas, M. Rowan-Robinson, R. Khardon, C. Alcock, and Y. Byun. "A Refined QSO Selection Method Using Diagnostics Tests: 663 QSO Candidates in the Large Magellanic Cloud." In *The Astrophysical Journal*, 747(2): 107, 2012.

[21] T. Muraveva, G. Clementini, C. Maceroni, C.J. Evans, M.I. Moretti, M.-R.L. Cioni, J.B. Marquette, V. Ripepi, R.de Grijs, M.A.T. Groenewegen, A.E. Piatti, and J.Th. van Loon. "EROS catalogue of eclipsing binary stars in the bar of the Large Magellanic Cloud." *Astronomy and Astrophysics Supplement Series*, 109: 447-469, 1995.

[22] D. Pelleg and A.W. Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters."*ICML*. 2000.

[23] I. Nun, P. Protopapas, B. Sim, M. Zhu, R. Dave, N. Castro, and K. Pichara. "FATS: Feature Analysis for Time Series 2015, arXiv:1506.00010

[24] J.W. Richards, D.L. Starr, Hk. Brink, A.A. Miller, J.S. Bloom, N.R. Butler, J.B. James, J.P. Long, and J. Rice. "Active learning to overcome sample selection bias: application to photometric variable star classification." In *The Astrophysical Journal*, 744(2): 192, 2012.

[25] M. Sviridenko. "A note on maximizing a submodular set function subject to a knapsack constraint." *Operations Research Letters*. 32.1 (2004): 41-43.

[26] S. Tong and D. Koller. "Support vector machine active learning with applications to text classification." TIn *he Journal of Machine Learning Research.*, 2: 45-66, 2002.

[27] J.A. Tyson. "Large synoptic survey telescope: overview." *Astronomical Telescopes and Instrumentation.International Society for Optics and Photonics*. 10-20, 2002.

[28] A. Udalski. "The optical gravitational lensing experiment. Real time data analysis systems in the OGLE-III survey." arXiv preprint astro-ph/0401123, 2004.

[29] M.K. Warmuth, J. Liao, G. Rtsch, et al. "Active learning with support vector machines in the drug discovery process." In *Journal of Chemical Information and Computer Sciences* 43(2): 667-673, 2003.

[30] E. Weinberger. "Correlated and uncorrelated fitness landscapes and how to tell the difference". In*Biological cybernetics*, 63(5): 325-336, 1990.

[31] J.J. Yang, X. Wang, P. Protopapas, and L. Bornn. "Fast and optimal nonparametric sequential design for astronomical observations." arXiv preprint arXiv:1501.02467, 2015.

[32] T. Zhang and F. Oles. "The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning*. 1191-1198, 2000.

[33] J. Zhu, H. Wang and E. Hovy. "Multi-criteria-based strategy to stop active learning for data annotation." In *Proceedings of the 22nd International Conference on Computational Linguistics*. Volume 1. Association for Computational Linguistics. 1129-1136, 2008

[34] X. Zhu, J. Lafferty, and Z. Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*. 2003.