Batch Mode Active Learning and Its Application to Medical Image Classification

Steven C. H. Hoi[†] CHHOI@CSE.CUHK.EDU.HK
Rong Jin[‡] RONGJIN@CSE.MSU.EDU
Jianke Zhu[†] JKZHU@CSE.CUHK.EDU.HK
Michael R. Lyu[†] LYU@CSE.CUHK.EDU.HK

[†]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[‡]Department of Computer Science and Engineering, Michigan State University, USA

Abstract

The goal of active learning is to select the most informative examples for manual labeling. Most of the previous studies in active learning have focused on selecting a *single* unlabeled example in each iteration. This could be inefficient since the classification model has to be retrained for every labeled example. In this paper, we present a framework for "**batch mode active learning**" that applies the Fisher information matrix to select a number of informative examples simultaneously. The key computational challenge is how to efficiently identify the subset of unlabeled examples that can result in the largest reduction in the Fisher information. To resolve this challenge, we propose an efficient greedy algorithm that is based on the property of submodular functions. Our empirical studies with five UCI datasets and one real-world medical image classification show that the proposed batch mode active learning algorithm is more effective than the state-of-the-art algorithms for active learning.

1. Introduction

Data classification has been an active research topic in machine learning in recent years. One of the prerequisites for any data classification scheme is the labeled examples. To reduce the effort involved in acquiring labeled examples, a number of active learning methods (Fine et al., 2002; Freund et al., 1997; Graepel & Herbrich, 2000; Seung et al., 1992; Campbell et al., 2000; Schohn & Cohn, 2000; Tong & Koller, 2000) have been developed in order to identify the examples that are most informative to the current classification model. In the past, active learning has been successfully employed in a number of applications, including text categorization (McCallum & Nigam, 1998; Roy & McCallum, 2001; Tong & Koller, 2000), computer vision (Luo et al., 2004), and information retrieval (Shen & Zhai, 2005).

Despite extensive studies of active learning, one of the main problems with most of the existing approaches is that only a *single* example is selected for manual labeling. As a result, the classification model has to be retrained after each labeled example is solicited. In this paper, we propose a novel active learning algorithm that is able to select a *batch* of unlabeled examples simultaneously. A simple strategy toward achieving batch mode active learning is to select the top k most informative examples. The problem with such an approach is that some of the selected examples could be similar, or even identical, to each other, and therefore do not provide additional information for model updating. Hence, the key of *batch mode active learning* is that, on the one hand, all the selected examples should be informative, and on the other hand, each selected example should be different from the others and should provide unique information.

To this end, we propose a framework of batch mode active learning that applies the Fisher information matrix to measure the overall informativeness for a set of unlabeled examples. The main computational challenge with the proposed framework is how to efficiently identify the subset of examples that are overall the most informative to the current classification model. To address the computational difficulty, we suggest an efficient greedy algorithm that is based on the properties of submodular functions.

To evaluate the effectiveness of the proposed active learning algorithms, we apply them to the task of med-

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

ical image classification. Recently, the application of machine learning techniques to medical image retrieval and analysis has received more and more attention (Lehmann et al., 2005). Due to the rapid development of computer technology, it is becoming more and more convenient to acquire, digitally store and transfer medical imagery. Nowadays, many hospitals need to manage several tera-bytes of medical image data each year (Müller et al., 2004). Therefore, categorization of medical images is becoming imperative for a variety of medical systems, especially in the application of digital radiology such as computer-aided diagnosis or case-based reasoning (Lehmann et al., 2005). Since annotating medical images can only be done by doctors with special expertise, acquiring labeled examples in medical image classification is usually substantially more expensive than in other classification problems. This special feature of medical image classification makes it more suitable for active learning.

The rest of this paper is organized as follows. Section 2 reviews the related work on medical image categorization and active learning algorithms. Section 3 presents the general framework for batch mode active learning. Section 4 describes two efficient algorithms for identifying a batch of unlabeled examples that are most informative to the classification model. Section 5 presents the empirical evaluation of our active learning methods in medical image categorization. Section 6 sets out our conclusion.

2. Related Work

Although image categorization in general is not new to researchers in computer science, only a few studies have been devoted to the medical domain. Only in recent years have researchers started to pay more attention to automatic categorization of medical images (Lehmann et al., 2005). Many classification algorithms have been applied to medical image categorization, including the large margin classifiers, decision trees, and neural networks. Among them, the large margin classifiers, such as support vector machines (SVM) and kernel logistic regression (KLR), appear to be most effective (Lehmann et al., 2005). One of the prerequisites for any classification scheme is the availability of labeled examples. Acquiring labeling information is usually a costly task. This is particularly true for medical image categorization since medical images can only be annotated and labeled by doctors. Hence, it is critical to reduce the labeling efforts for medical image categorization. To address this problem, we apply active learning techniques that select only the most informative medical images for doctors to label. It is worth noting that another approach to reduce the labeling efforts is the semi-supervised learning methods (Seeger, 2001), which learns a classification model from a mixture of labeled and unlabeled data.

Active learning, or called pool-based active learning, has been extensively studied in machine learning for several years (McCallum & Nigam, 1998; Roy & McCallum, 2001). Most active learning algorithms are conducted in an iterative fashion. In each iteration, the example with the highest classification uncertainty is chosen for manual labeling. Then, the classification model is retrained with the additional labeled example. The steps of training a classification model and soliciting a labeled example are iterated alternatively until most of the examples can be classified with reasonably high confidence. One of the key issues in active learning is how to measure the classification uncertainty of the unlabeled examples. In several recent studies (Fine et al., 2002; Freund et al., 1997; Graepel & Herbrich, 2000; Seung et al., 1992), a number of distinct classification models are first generated. Then, the classification uncertainty of a test example is measured by the amount of disagreement among the ensemble of classification models in predicting the labels for the test example. Another group of approaches measures the classification uncertainty of a test example by how far the example is away from the classification boundary (i.e., classification margin) (Campbell et al., 2000; Schohn & Cohn, 2000; Tong & Koller, 2000). One of the most well-known approaches within this group is *Support Vector Machine Active Learning*, developed by Tong and Koller (Tong & Koller, 2000). Due to its popularity and success in previous studies, we use it as the baseline approach in our study.

One of the main problems with most existing active learning algorithms is that only a *single* example is selected for labeling each time. As a result, the classification model has to be retrained after each labeled example is solicited. In this paper, we focus on the batch mode active learning that selects a *batch* of unlabeled examples in each iteration. A simple strategy is to choose the top k most uncertain examples. However, it is likely that some of the most uncertain examples are strongly correlated and therefore will provide similar information to the classification model. In general, the challenge in choosing a batch of unlabeled examples is twofold: on the one hand, the examples in the selected batch should be informative to the classification model; on the other hand, the examples should be diverse enough such that information provided by different examples does not overlap. To address this challenge, we employ the Fisher information matrix

as the measurement of model uncertainty, and choose the set of examples that efficiently reduces the Fisher information.

3. A Framework of Batch Mode Active Learning

In this section, we describe the framework for batch mode active learning that is based on the Fisher information matrix. We choose the logistic regression model as the underlying classification model because of its simplicity and probabilistic nature. To facilitate our discussion, we start with the linear classification model, followed by the extension to the nonlinear classification model using the kernel trick.

The theoretical foundation of our batch mode active learning is based on the work of (Zhang & Oles, 2000), in which the authors presented a framework of active learning based on the maximization of the Fisher information matrix. Given that the Fisher information matrix represents the overall uncertainty of a classification model, our goal is to search for a set of examples that can most efficiently reduce the Fisher information matrix. More specifically, this goal can be formulated into the following optimization problem:

Let $p(\mathbf{x})$ be the distribution of all unlabeled examples, and $q(\mathbf{x})$ be the distribution of unlabeled examples that are chosen for manual labeling. Let α denote the parameters of the classification model. Let $I_p(\alpha)$ and $I_q(\alpha)$ denote the Fisher information matrix of the classification model for the distribution $p(\mathbf{x})$ and $q(\mathbf{x})$, respectively. Then, the set of examples that can most efficiently reduce the uncertainty of classification model is found by minimizing the ratio between the two Fisher information matrices $I_p(\alpha)$ and $I_q(\alpha)$, i.e.,

$$q^* = \arg\min_q \operatorname{tr}(I_q(\alpha)^{-1} I_p(\alpha)) \quad (1)$$

For logistic regression models, the Fisher information matrix $I_q(\alpha)$ is obtained by:

$$\begin{aligned} & I_q(\alpha) \\ &= -\int q(\mathbf{x}) \sum_{y=\pm 1} p(y|\mathbf{x}) \frac{\partial^2}{\partial \alpha^2} \log p(y|\mathbf{x}) d\mathbf{x} \\ &= \int \frac{1}{1+\exp(\alpha^T \mathbf{x})} \frac{1}{1+\exp(-\alpha^T \mathbf{x})} \mathbf{x}\mathbf{x}^T q(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (2)$$

In order to estimate the optimal distribution $q(\mathbf{x})$, we replace the integration in the above equation with a summation over the unlabeled data and the selected examples. Let $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the unlabeled data, and $S = (\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_k^s)$ be the subset of selected examples, where k is the number of examples to be selected. We can now rewrite the above expression for Fisher information matrices I_p and I_q as:

$$I_p(\hat{\alpha}) = \frac{1}{n} \sum_{\mathbf{x} \in D} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \mathbf{x}\mathbf{x}^T + \delta I_d$$

$$I_q(S, \hat{\alpha}) = \frac{1}{k} \sum_{\mathbf{x} \in S} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \mathbf{x}\mathbf{x}^T + \delta I_d$$

where

$$\pi(\mathbf{x}) = p(-|\mathbf{x}) = \frac{1}{1+\exp(\hat{\alpha}^T \mathbf{x})} \quad (3)$$

In the above, $\hat{\alpha}$ stands for the classification model that is estimated from the labeled examples. I_d is the identity matrix of size $d \times d$. $\delta \ll 1$ is the smoothing parameter. δI_d is added to the estimation of $I_p(\hat{\alpha})$ and $I_q(S, \hat{\alpha})$ to prevent them from being singular matrices. Hence, the final optimization problem for batch mode active learning is formulated as follows:

$$S^* = \arg \min_{S \subseteq D \wedge |S| = k} \operatorname{tr}(I_q(S, \hat{\alpha})^{-1} I_p(\alpha)) \quad (4)$$

To extend the above analysis to the nonlinear classification model, we follow the idea of imported vector machine (Zhu & Hastie, 2001) by introducing a kernel function $K(\mathbf{x}', \mathbf{x})$ and rewriting the logistic regression model as:

$$p(y|\mathbf{x}) = \frac{1}{1+\exp(-yK(\mathbf{w}, \mathbf{x}))}$$

According to the representer theorem, $\phi(\mathbf{w})$ could be written as a linear combination of $\phi(\mathbf{x})$ for the labeled examples $\mathbf{x}$, i.e.,

$$\phi(\mathbf{w}) = \sum_{\mathbf{x} \in L} \theta(\mathbf{x}) \phi(\mathbf{x})$$

where $\theta(\mathbf{x})$ is the combination weight for labeled example $\mathbf{x}$. $L = ((y_1, \mathbf{x}_1^L), (y_2, \mathbf{x}_2^L), \ldots, (y_m, \mathbf{x}_m^L))$ stands for the set of labeled examples, where m is the number of labeled examples. Using the result of representer theorem, we have $K(\mathbf{w}, \mathbf{x})$ and $p(y|\mathbf{x})$ rewritten as:

$$\begin{aligned} K(\mathbf{w}, \mathbf{x}) &= \sum_{\mathbf{x}' \in L} \theta(\mathbf{x}') K(\mathbf{x}', \mathbf{x}) \\ p(y|\mathbf{x}) &= \frac{1}{1+\exp\left(-y \sum_{\mathbf{x}' \in L} \theta(\mathbf{x}') K(\mathbf{x}', \mathbf{x})\right)} \end{aligned}$$

Thus, by treating $(K(\mathbf{x}_1^L, \mathbf{x}), K(\mathbf{x}_2^L, \mathbf{x}), \ldots, K(\mathbf{x}_m^L, \mathbf{x}))$ as the new representation for the unlabeled example $\mathbf{x}$, we can directly apply the result for the linear logistic regression model to the nonlinear case.

4. Efficient Algorithms for Batch Mode Active Learning

The challenge with solving the optimization problem in Eqn. (4) is that the number of candidate sets for S is exponential in the number of unlabeled examples n. As a result, it is computationally prohibitive when the number of unlabeled examples is large. In order to resolve the difficulty with the combinatorial optimization, we present a greedy algorithm that is based on the idea of submodular function.

The key idea of this approach is to explore the general theorem about submodular functions in (Nemhauser et al., 1978): consider the optimization problem that searches for a subset S with k elements to maximize a set function $f(S)$, i.e.,

$$\max_{|S|=k} \quad f(S)$$

If $f(S)$ is 1) a nondecreasing submodular function, and 2) $f(\emptyset) = 0$, then the greedy algorithm will guarantee a performance $(1 - 1/e)f(S^*)$, where $S^* = \arg\max_{|S|=k} f(S)$ is the optimal subset. Based on this theorem, when a set function $f(S)$ satisfies the two conditions, namely nondescreasing submodular function and $f(\emptyset) = 0$, the subset S that maximizes $f(S)$ can be well approximated by the solution obtained by the greedy algorithm.

In order to utilize the above theorem, the key is to approximate the objective function in Eqn. (4) by a submodular function. To this end, we simplify the objective function as follows:

$$\begin{aligned}
&\mathrm{tr}(I_q^{-1}(S)I_p) \\
&= \mathrm{tr}\left(I_q^{-1}(S)\left(\begin{array}{l} I_q(S)\frac{k}{n} + \frac{n-k}{n}\delta I + \\ \frac{1}{n}\sum_{\mathbf{x}\notin\mathcal{S}} \pi(\mathbf{x})(1-\pi(\mathbf{x}))\mathbf{x}\mathbf{x}^\top \end{array}\right)\right) \\
&= \frac{k}{n} + \delta\frac{n-k}{n}\mathrm{tr}(I_q^{-1}(S)) \\
&\quad + \frac{1}{n}\sum_{\mathbf{x}\notin S} \pi(\mathbf{x})(1-\pi(\mathbf{x}))\mathbf{x}^\top I_q^{-1}(S)\mathbf{x}
\end{aligned}$$

We ignore the second term in the above expression, i.e., $\delta(n-k)\mathrm{tr}(I_q^{-1}(S))/n$, and only focus on the last term, i.e., $\sum_{\mathbf{x}\notin\mathcal{S}} \mathbf{x}^\top I_q^{-1}(S)\mathbf{x}$. This is because the second term is proportional to the smoothing parameter δ that is usually set to be small. To further simplify the computation, we approximate the term $\mathbf{x}^\top I_q^{-1}(S)\mathbf{x}$ as follows:

Let $\{(\lambda_k, \mathbf{v}_k)\}_{k=1}^d$ be the eigenvectors of matrix $I_q(S)$. Then, for any $\mathbf{x}$, we have

$$\begin{aligned}
\mathbf{x}^T I_q^{-1}(S)\mathbf{x} &= \sum_{k=1}^d \lambda_k^{-1}(\mathbf{x}^T\mathbf{v}_k)^2 \\
&\approx \frac{\|\mathbf{x}\|_2^2}{\sum_{k=1}^d \lambda_k(\mathbf{x}^\top\mathbf{v}_k)^2/\|\mathbf{x}\|_2^2} = \frac{(\sum_{i=1}^d x_i^2)^2}{\mathbf{x}^\top I_q(S)\mathbf{x}}
\end{aligned}$$

In the above, we approximate the harmonic mean of the eigenvalues λ_is by their arithmetic mean, i.e., $\left(\sum_{i=1}^d \lambda_i^{-1} p_i\right)^{-1} \approx \sum_{i=1}^d \lambda_i p_i$ where $p_i = (\mathbf{x}^T\mathbf{v}_i)^2 / \left(\sum_{i=1}^d (\mathbf{x}^T\mathbf{v}_i)\right)^2 = (\mathbf{x}^T\mathbf{v}_i)^2/\|\mathbf{x}\|_2^2$ is a PDF. Note that this approximation will make the optimal solution more stable than the original objective function. This is because $\mathrm{tr}(I_q^{-1}(S)I_p)$ is proportional to λ_i^{-1} and therefore is sensitive to the small eigenvalues of I_q while the approximate one does not.

By assuming that each example $\mathbf{x}$ is normalized as 1, namely $\|\mathbf{x}\|_2^2 = 1$, we have

$$\begin{aligned}
&\sum_{\mathbf{x}\notin S} \pi(\mathbf{x})(1-\pi(\mathbf{x}))\mathbf{x}^\top I_q^{-1}(S)\mathbf{x} \\
&\approx \sum_{\mathbf{x}\notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\mathbf{x}^\top I_q(S)\mathbf{x}} \\
&= \sum_{\mathbf{x}\notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))k}{\delta + \sum_{\mathbf{x}'\in S} \pi(\mathbf{x}')(1-\pi(\mathbf{x}'))(\mathbf{x}^\top\mathbf{x}')^2}
\end{aligned}$$

Hence, the entire optimization problem in Eqn. (4) is simplified as follows:

$$\min_{|S|=k\wedge S\subseteq D} \sum_{\mathbf{x}\notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\delta + \sum_{\mathbf{x}'\in S} \pi(\mathbf{x}')(1-\pi(\mathbf{x}'))(\mathbf{x}^\top\mathbf{x}')^2} \quad (5)$$

In order to explore the theorem about submodular functions described in (Nemhauser et al., 1978), we define the set function $f(S)$ as follows:

$$\begin{aligned}
f(S) &= \frac{1}{\delta}\sum_{\mathbf{x}\in D} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \quad (6) \\
&\quad - \sum_{\mathbf{x}\notin S} \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\delta + \sum_{\mathbf{x}'\in S} \pi(\mathbf{x}')(1-\pi(\mathbf{x}'))(\mathbf{x}^\top\mathbf{x}')^2}
\end{aligned}$$

Evidently, the problem in Eqn. (5) is equivalent to the following optimization problem:

$$\max_{|S|=k\wedge S\subseteq D} \quad f(S) \quad (7)$$

It is easy to see that $f(\emptyset) = 0$. It is also not difficult to show that $f(S)$ is a nondecreasing submodular function. The detailed proof can be found in the Appendix. Hence, the set function $f(S)$ satisfies the two

- **Initialize** $S = \emptyset$
- **For** $i = 1, 2, \ldots, k$
  - Compute $\mathbf{x}^* = \arg\max_{\mathbf{x} \notin \mathcal{S}} f(S \cup \mathbf{x}) - f(S)$
  - Set $S = S \cup \mathbf{x}^*$

Figure 1. A greedy algorithm for $\arg\max_{|S|=k} f(S)$

conditions of the theorem about submodular functions, and the result of the theorem can be applied directly to the problem in (7). More specifically, the value of the subset found by the greedy algorithm is no less than $1 - 1/e$ of the value of the true optimal subset. In Figure 1, we summarize the greedy algorithm that solves the optimization problem in (7).

Remark. To see what type of examples will be chosen by the greedy algorithm, we analyze the difference between $f(S \cup \mathbf{x})$ and $f(S)$, which is written as follows:

$$\begin{aligned} & f(S \cup \mathbf{x}) - f(S) \\ = \; & g(\mathbf{x}, S) + \sum_{\mathbf{x}' \notin (S \cup \mathbf{x})} g(\mathbf{x}', S) g(\mathbf{x}, S \cup \mathbf{x}) (\mathbf{x}^\top \mathbf{x}')^2 \end{aligned}$$

where function $g(\mathbf{x}, S)$ is defined as

$$g(\mathbf{x}, S) = \frac{\pi(\mathbf{x})(1 - \pi(\mathbf{x}))}{\delta + \sum_{\mathbf{x}' \in S} \pi(\mathbf{x}')(1 - \pi(\mathbf{x}'))(\mathbf{x}^\top \mathbf{x}')^2} \quad (8)$$

Based on the above expressions, we can draw the following observations:

- $f(S \cup \mathbf{x}) - f(S) \propto \pi(\mathbf{x})(1 - \pi(\mathbf{x}))$. This indicates that examples with large classification uncertainty are more likely to be selected than examples with small classification uncertainty.
- The first term in $f(S \cup \mathbf{x}) - f(S)$ is inverse to $\sum_{\mathbf{x}' \in S} \pi(\mathbf{x}')(1 - \pi(\mathbf{x}'))(\mathbf{x}^\top \mathbf{x})^2$. This indicates that the optimal choice of example $\mathbf{x}$ should **not** be similar to examples in $\mathcal{S}$, i.e., the set of selected instances.
- The second term in $f(S \cup \mathbf{x}) - f(S)$ is proportional to $(\mathbf{x}' \mathbf{x})^2$ for all examples $\mathbf{x}' \notin S$. This indicates that the optimal choice of example $\mathbf{x}$ should be similar to the unselected examples.

In summary, the selected examples will have the following three properties: 1) uncertain to the current classification model, 2) dissimilar to the other selected examples, and 3) similar to most of the unselected examples. Clearly, these three properties are the desirable properties for batch mode active learning.

5. Experimental Result

In this section, we report our empirical study of batch mode active learning in the application to medical image categorization.

Table 1. List of UCI machine learning datasets.

Dataset	#Instances	#Features
Australian	690	14
Breast-cancer	683	10
Heart	270	13
Ionosphere	351	34
Sonar	208	60

Table 2. List of medical image categories.

Category	Category Info	#Instances
Cat-1	cranium, MUS	336
Cat-2	cervical spine, MUS	215
Cat-3	thoracic spine, MUS	102
Cat-4	lumbar spine, MUS	225
Cat-5	hand, MUS	576
Cat-6	radio carpal joint, MUS	77
Cat-7	elbow, MUS	69
Cat-8	shoulder, MUS	108
Cat-9	chest, bones, MUS	93
Cat-10	abdomen, GAS	152
Cat-11	pelvis, MUS	217
Cat-12	foot, MUS	205
Cat-13	ankle joint, MUS	137
Cat-14	knee, MUS	194
Cat-15	hip, MUS	79

MUS:"musculoskeletal system", GAS:"gastrointestinal system".

5.1. Experimental Testbeds

To examine the effectiveness of our active learning algorithm, we first evaluate the performance of the proposed batch active learning algorithm on five datasets from the UCI machine learning repository[1]. Table 1 shows the list of the datasets used in our experiment.

We then evaluate the proposed batch mode active learning algorithm on medical image classification. The medical image dataset is formed by randomly selecting 2,785 medical images from the ImageCLEF (Lehmann et al., 2005) that belong to 15 different categories. Table 2 gives the details of the medical image testbed. Each image is represented by 2560 visual features that are extracted by the Gabor wavelet transform. To represent the visual characteristics of medical images, the Gabor wavelet filters (Liu & Wechsler, 2002) are emplyed to extract the texture features of medical images.

5.2. Empirical Evaluation

Since medical image categorization is a classification problem, we adopt the classification $F1$ performance as the evaluation metric. The $F1$ metric is defined as $F1 = 2 * p * r/(p + r)$, the harmonic mean of precision

[1] www.ics.uci.edu/ mlearn/MLRepository.html

p and recall r of classification. Since the $F1$ metric takes into account both the precision and the recall of classification, it is usually preferred over other metrics.

In our experiments, two large margin classifiers, i.e., the kernel logistic regressions (KLR) (Zhu & Hastie, 2001) and the support vector machines (SVM) (Burges, 1998), are employed as the basis classifiers. Two active learning algorithms are employed as the baseline models in our studies. The first baseline model is the kernel logistic regression active learning algorithm that measures the classification uncertainty based on the entropy of the distribution $p(y|\mathbf{x})$. The examples with the largest entropy are selected for manual labeling. We refer to this baseline model as the logistic regression active learning model, or **KLR-AL** for short. The second reference model is based on support vector machine active learning (Tong & Koller, 2000). In this method, the classification uncertainty of an example $\mathbf{x}$ is determined by its distance from the decision boundary $\mathbf{w}^T\mathbf{x} + b = 0$. The unlabeled examples with the smallest distance are selected for labeling. We refer to this approach as SVM active learning, or **SVM-AL** for short.

To evaluate the performance of the competing active learning algorithms, we first randomly pick l training samples from the dataset for each category consisting of an equal number of negative and positive examples. We then train both SVM and KLR classifiers using the l labeled examples, respectively. Based on these two initially trained models, additional s (referred to as the "batch size") unlabeled examples are chosen for manual labeling for each active learning method. To see the effects of the selected examples on the classification models, we also train the two reference models by randomly selecting s examples for manually labeling, which are referred to as **SVM-Rand** and **KLR-Rand**, respectively. For performance comparison, every experiment is carried out 20 times, and the averaged classification $F1$ with their standard errors are calculated and used for final evaluation.

5.3. Experimental Results

Table 3 summarizes the experimental results of the five UCI datasets for the proposed batch mode active learning algorithm as well as the two baseline approaches for active learning. Both the number of initially labeled examples l and the number of selected examples s for each iteration are set to be 10. Due to the space limitation, we only presented the results of the first two iterations. Compared to the two reference models using randomly selected examples, i.e., SVM-Rand and KLR-Rand, all the three active learning algorithms are able to achieve noticeable improvement in $F1$ across all five UCI datasets. We also observed that the improvement made by the active learning algorithms in the second iteration is considerably smaller than that of the first iteration. In fact, after the third iteration, the improvement made by the active learning algorithms starts to diminish.

Comparing to the two baseline active learning algorithms, we observe that the proposed approach performs significantly ($p < 0.05$) better over the dataset "Australian", "Ionosphere", and "Sonar", according to the *student-t test*. We further examine the performance of the proposed approach by varying the batch size from 10 to 50. Fig. 2 shows the experimental results of the three active learning methods using different batch sizes. Again, we observe that the batch mode active learning method consistently outperforms the other methods across different batch sizes.

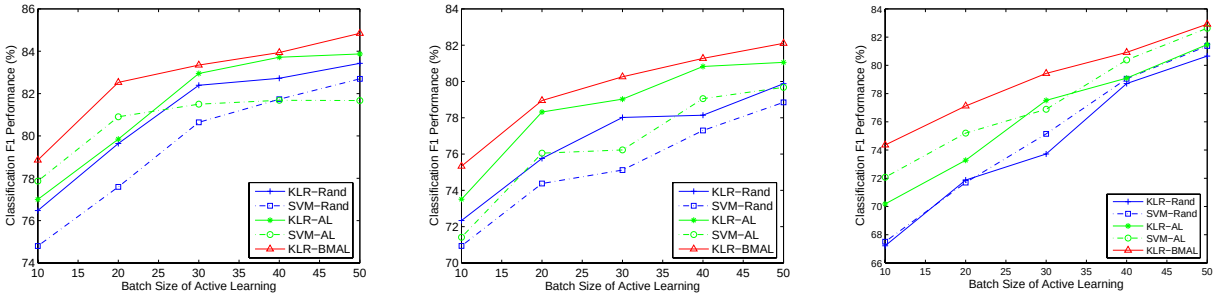
Table 4 summarizes the experimental results of the first two iterations for the medical image dataset. 40 labeled examples are used for initially training, and 20 examples are selected for each iteration of active learning. Similar to the UCI datasets, the three active learning algorithms perform considerably better than the two reference models across all the categories. The most noticeable case is category 3, where $F1$ is improved from around 40% to about 50% by the active learning algorithms. Furthermore, the comparison between the batch mode active learning algorithm and the two non-batch active learning algorithms revealed that the proposed algorithm for batch mode active learning always improves the classification performance. For a number of categories, including category 3, 10, 12, and 15, the improvement in the $F1$ measurement is statistically significant ($p < 0.05$) according to the *student-t test*. Similar improvements were also observed for different batch sizes. We did not report those results due to the space limitation.

6. Conclusion

This paper presented a general framework for batch mode active learning. Unlike the traditional active learning that focuses on selecting a single example in each iteration, the batch mode active learning allows multiple examples to be selected for manual labeling. We use the Fisher information matrix for the measurement of model uncertainty and choose the set of examples that will effectively reduce the Fisher information. In order to solve the related optimization problem, we proposed an efficient greedy algorithm that approximates the objective function by a submodular function. Empirical studies with five UCI

Table 3. Evaluation of classification F1 performance on the UCI datasets.

Dataset	Active Learning Iteration-1					Active Learning Iteration-2				
	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL
Australian	74.80 ±1.97	76.48 ±2.16	77.86 ±0.84	77.00 ±1.14	**78.86** ±1.00	79.29 ±1.30	80.89 ±1.29	80.73 ±0.93	81.43 ±0.89	**83.49** ±0.36
Breast	96.34 ±0.37	96.10 ±0.33	96.80 ±0.20	97.05 ±0.02	**97.67** ±0.06	96.80 ±0.23	96.26 ±0.55	97.52 ±0.07	97.71 ±0.06	**97.81** ±0.03
Heart	70.94 ±1.29	72.34 ±1.46	71.41 ±2.39	73.51 ±1.80	**75.33** ±1.26	76.76 ±0.70	77.84 ±0.78	76.92 ±0.91	78.78 ±1.12	**79.53** ±0.59
Ionosphere	88.58 ±0.83	88.78 ±0.81	89.05 ±1.12	89.66 ±1.10	**92.39** ±0.69	90.45 ±0.59	90.60 ±0.61	93.42 ±0.51	93.71 ±0.49	**94.26** ±0.55
Sonar	67.51 ±1.57	67.22 ±1.49	72.07 ±0.84	70.18 ±1.28	**74.36** ±0.43	73.80 ±0.81	73.33 ±0.97	75.11 ±0.87	74.80 ±0.78	**77.49** ±0.45



(a) Australian (b) Heart (c) Sonar

Figure 2. Evaluation of classification F1 performance on the UCI datasets with different batch sizes.

datasets and one medical image dataset demonstrated that the proposed batch mode active learning algorithm is more effective than the margin-based active learning approaches, which have been the dominant methods for active learning.

Acknowledgements

The work described in this paper was fully supported by two grants, one from the Shun Hing Institute of Advanced Engineering, and the other from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4205/04E).

Appendix

Theorem 1 *The set function $f(S)$ in Eq. (6) is a nondecreasing submodular function.*

Proof. To prove that the set function $f(S)$ is a submodular function, we use the sufficient and necessary condition for submodular functions (Parker, 1988), i.e., for any two sets $A \subseteq B$, for any element $\mathbf{x} \notin B$, $f(S)$ is a submodular function if and only if the following condition holds:

$$f(A \cup \mathbf{x}) - f(A) \geq f(B \cup \mathbf{x}) - f(B)$$

In order to show the above property, we compute the difference $f(S \cup \mathbf{x}) - f(S)$ for $\mathbf{x} \notin S$, i.e.,

$$\begin{aligned} & f(S \cup \mathbf{x}) - f(S) \\ = \; & g(\mathbf{x}, S) + \sum_{\mathbf{x}' \notin (S \cup \mathbf{x})} g(\mathbf{x}', S) g(\mathbf{x}, S \cup \mathbf{x}) (\mathbf{x}^\top \mathbf{x}')^2 \end{aligned}$$

where the function $g(\mathbf{x}, S)$ is already defined in Eqn. (8). First, according to the definition of function $g(\mathbf{x}, S)$ in Eqn. (8), $g(\mathbf{x}, S) \geq 0$ for any $\mathbf{x}$ and $\mathcal{S}$. Thus we have $f(S \cup \mathbf{x}) \geq f(S)$; therefore, $f(S)$ is a nondescreasing function. Second, as indicated by the above expression, the difference $f(S \cup \mathbf{x}) - f(S)$ is a monotonically decreasing function. As a result, we have $f(A \cup \mathbf{x}) - f(A) \geq f(B \cup \mathbf{x}) - f(B)$ when $A \subseteq B$. In conclusion, the function $f(S)$ is a nondecreasing submodular function.

References

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Campbell, C., Cristianini, N., & Smola, A. J. (2000). Query learning with large margin classifiers. *Int. Conf. on Machine Learning (ICML)* (pp. 111–118).

Fine, S., Gilad-Bachrach, R., & Shamir, E. (2002). Query by committee, linear separation and random walks. *Theor. Comput. Sci.*, *284*, 25–51.

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Mach. Learn.*, *28*, 133–168.

Graepel, T., & Herbrich, R. (2000). The kernel gibbs sampler. *NIPS 13* (pp. 514–520).

Table 4. Evaluation of classification F1 performance on the medical image datasets.

Dataset	Active Learning Iteration-1					Active Learning Iteration-2				
	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL
Cat-1	90.42 ±0.64	91.49 ±0.52	94.33 ±0.39	95.07 ±0.25	**95.63** ±0.22	91.72 ±0.63	93.56 ±0.42	96.44 ±0.25	96.96 ±0.21	**97.32** ±0.10
Cat-2	66.68 ±1.94	70.68 ±1.92	74.92 ±1.48	77.54 ±1.29	**79.40** ±1.52	73.96 ±1.39	78.93 ±1.13	84.27 ±0.52	86.17 ±0.51	**86.72** ±0.52
Cat-3	41.33 ±1.37	44.11 ±1.40	50.51 ±1.93	54.94 ±1.77	**56.81** ±1.72	50.08 ±1.69	56.72 ±1.45	68.67 ±1.34	76.52 ±0.68	**77.06** ±0.91
Cat-4	68.09 ±1.70	69.34 ±1.65	77.08 ±1.40	79.22 ±1.40	**80.83** ±1.23	74.88 ±1.42	78.27 ±1.11	88.00 ±0.87	89.51 ±0.74	**89.80** ±0.72
Cat-5	62.10 ±0.82	64.12 ±0.66	63.61 ±0.65	64.52 ±0.68	**65.76** ±0.71	66.27 ±0.87	67.79 ±0.75	68.18 ±0.61	69.64 ±0.74	**69.73** ±0.55
Cat-6	49.47 ±2.19	48.97 ±2.72	54.80 ±1.79	56.66 ±2.29	**60.20** ±2.03	56.10 ±2.40	55.02 ±2.36	70.49 ±0.92	71.17 ±1.09	**71.99** ±1.19
Cat-7	33.26 ±1.43	35.04 ±1.51	34.73 ±1.63	33.92 ±1.01	**34.12** ±0.98	41.13 ±1.54	42.75 ±1.48	48.27 ±1.71	49.95 ±1.47	**51.50** ±1.43
Cat-8	30.85 ±0.82	32.79 ±0.91	32.85 ±1.20	36.87 ±1.31	**37.46** ±1.46	37.17 ±1.05	42.54 ±1.36	45.76 ±1.37	51.69 ±1.71	**54.51** ±1.42
Cat-9	25.72 ±0.98	26.09 ±1.01	29.50 ±1.16	29.28 ±1.29	**30.17** ±1.32	32.70 ±1.05	34.07 ±1.13	42.06 ±1.53	43.27 ±1.89	**44.76** ±1.62
Cat-10	47.19 ±1.75	47.72 ±1.38	51.23 ±2.18	51.03 ±1.68	**53.27** ±1.56	57.60 ±1.66	56.45 ±1.53	57.88 ±2.46	62.62 ±1.70	**63.64** ±2.21
Cat-11	74.40 ±1.82	79.65 ±1.85	80.81 ±1.51	83.99 ±1.15	**85.69** ±0.75	79.43 ±1.28	84.53 ±0.86	87.54 ±0.55	89.72 ±0.40	**90.21** ±0.42
Cat-12	34.11 ±1.05	35.81 ±1.03	35.91 ±0.84	36.37 ±0.93	**37.54** ±0.92	38.93 ±1.04	40.30 ±0.95	43.57 ±1.20	43.79 ±1.14	**45.59** ±1.23
Cat-13	65.00 ±1.61	64.11 ±1.94	71.97 ±0.88	74.89 ±1.34	**76.37** ±1.17	71.14 ±0.70	73.01 ±1.07	78.66 ±0.92	84.40 ±0.66	**85.24** ±0.39
Cat-14	60.06 ±1.19	60.50 ±1.30	66.45 ±1.22	68.96 ±1.18	**69.82** ±1.06	66.42 ±0.97	67.46 ±0.94	74.84 ±0.54	77.98 ±0.72	**78.71** ±0.47
Cat-15	30.90 ±1.43	32.58 ±1.57	33.96 ±2.36	33.73 ±2.07	**34.37** ±2.20	40.69 ±2.17	43.88 ±2.38	52.37 ±2.25	53.63 ±2.32	**55.31** ±2.30

Lehmann, T. M., Güld, G. O., T.Deselaers, Keysers, D., Schubert, H., Spitzer, K., Ney, H., & Wein, B. B. (2005). Automatic categorization of medical images for content-based retrieval and data mining. *Computerized Medical Imaging and Graphics*, *29*, 143–155.

Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, *11*, 467–476.

Luo, T., Kramer, K., Samson, S., & Remsen, A. (2004). Active learning to recognize multiple types of plankton (pp. III: 478–481.).

McCallum, A. K., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. *Int. Conf. on Machine Learning* (pp. 350–358).

Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Int J Med Inform*, *73*, 1–23.

Nemhauser, G., Wolsey, L., & Fisher, M. (1978). An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, *14*, 265–294.

Parker, R. G. (1988). *Discrete optimization*. Academic Press.

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *18th ICML* (pp. 441–448).

Schohn, G., & Cohn, D. (2000). Less is more: Active learning with support vector machines. *17th ICML* (pp. 839–846).

Seeger, M. (2001). *Learning with labeled and unlabeled data* (Technical Report). University of Edinburgh.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Computational Learning Theory* (pp. 287–294).

Shen, X., & Zhai, C. (2005). Active feedback in ad hoc information retrieval. *Proc. ACM SIGIR'05* (pp. 59–66). Salvador, Brazil.

Tong, S., & Koller, D. (2000). Support vector machine active learning with applications to text classification. *17th ICML* (pp. 999–1006). Stanford, US.

Zhang, T., & Oles, F. J. (2000). A probability analysis on the value of unlabeled data for classification problems. *17th ICML* (pp. 1191–1198). Stanford, US.

Zhu, J., & Hastie, T. (2001). Kernel logistic regression and the import vector machine. *Advances in Neural Information Processing Systems 14* (pp. 1081–1088).