Scalable Histopathological Image Analysis via Active Learning

Yan Zhu[1], Shaoting Zhang[2], Wei Liu[3], and Dimitris N. Metaxas[1]

[1] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[2] Department of Computer Science, University of North Carolina at Charlotte, NC, USA
[3] IBM T.J. Watson Research Center, NY, USA

Abstract. Training an effective and scalable system for medical image analysis usually requires a large amount of labeled data, which incurs a tremendous annotation burden for pathologists. Recent progress in active learning can alleviate this issue, leading to a great reduction on the labeling cost without sacrificing the predicting accuracy too much. However, most existing active learning methods disregard the "structured information" that may exist in medical images (*e.g.*, data from individual patients), and make a simplifying assumption that unlabeled data is independently and identically distributed. Both may not be suitable for real-world medical images. In this paper, we propose a novel batch-mode active learning method which explores and leverages such structured information in annotations of medical images to enforce diversity among the selected data, therefore maximizing the information gain. We formulate the active learning problem as an adaptive submodular function maximization problem subject to a partition matroid constraint, and further present an efficient greedy algorithm to achieve a good solution with a theoretically proven bound. We demonstrate the efficacy of our algorithm on thousands of histopathological images of breast microscopic tissues.

1 Introduction

Recent development of microscopical acquisition technology enables computerized analysis of histopathological images [9]. For example, in the context of breast cancer diagnosis, plenty of systems have been designed to conduct automatic and accurate analysis of high-resolution images digitized from tissue histopathology slides, where well-known machine learning and image processing techniques [12,3,4] have been exploited. Particularly, supervised learning models such as Support Vector Machines (SVMs) [13] have been extensively employed, because they are able to effectively bridge the so-called "semantic gap" between histopathological images and their diagnosis information [3,6,9]. To train an accurate prediction model under a supervised manner, it is usually necessary to require a large amount of labeled data, *e.g.*, manual annotations from domain experts or pathologists. However, acquiring sufficient high-quality annotations is a very expensive and tedious process. To alleviate this issue and reduce the labeling cost, active learning [14] has been suggested to intelligently select a small yet informative subset of the whole database, which requires only a few labeling operations from domain experts to build an accurate enough prediction model yet with a low training cost.

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 369–376, 2014.

Active learning has been widely investigated in the machine learning community, aiming for progress in both theoretical aspects, *e.g.*, sample complexity bounds [1], and approaching practical applications, *e.g.*, image [10] and text [15] classification and retrieval (the related work in active learning is briefly described below). However, for histopathological images, previous active learning methods have two main shortcomings: 1) Almost all of them assume that unlabeled data samples are *independently and identically distributed* (I.I.D.), which is not necessarily suitable for histopathological images. In fact, for each patient there are usually several images available which share common pathological characteristics, *e.g.*, images from different ROIs. Obviously, there are considerable correlations among such image samples. 2) Even if the I.I.D. property holds, previous active learning methods may disregard the structured information of histopathological images, *e.g.*, patient identity, which is easy to obtain but could be crucial for active learning to enforce diversity during sample selection.

In this work, we propose a novel batch mode active learning approach which is specifically designed for histopathological image analysis by leveraging structured information to enforce diversity during intelligent sample selection. We formulate the active learning problem (essentially the sample selection problem) as a constrained submodular optimization problem and present a greedy algorithm to efficiently solve it. Notably, we provide a theoretical bound characterizing the quality of the submodular active learning strategy, which guarantees that our proposed greedy algorithm approximates the optimal batch mode active learning strategy for the adaptive submodular function maximization problem with a partition matroid constraint. In practice, our active learning driven histopathological image analysis approach outperforms state-of-the-art methods to tackle histopathological image analysis. We perform experiments on a large database of histopathological images with high-dimensional features. The experimental results demonstrate the efficacy of our approach, which achieves 83% prediction accuracy with merely 100 labeled samples among more than two thousand images (*i.e.*, less than 5% training data). This accuracy is 11% higher than passive learning and 6% higher than state-of-the-art active learning methods.

Related Work in Active Learning. Active learning can be considered as a combinatorial optimization problem which is typically difficult to exactly solve, so a variety of heuristics have been resorted to. For example, a number of active learning algorithms relax the original combinatorial problem involving discrete constraints to a continuous optimization problem, and then employ regular convex or non-convex optimization techniques to solve the relaxed problem. These algorithms usually suffer from prohibitively high computational complexities, and the deviation from the solution of the relaxed problem to that of the original problem remains unknown. In contrast, some latest work casts active learning problem into a submodular set function maximization problem which is direct combinatorial optimization. While maximizing a submodular function appears NP-hard, a landmark result from Nemhauser *et al.* [5] certifies that a simple greedy optimization scheme is able to achieve the $(1 - \frac{1}{e})$-approximation for the cardinality constraint and the $(\frac{1}{p+1})$-approximation for p matroid constraints, respectively. Built on this theoretic finding, Chen and Krause [2] propose a nearly optimal batch mode active learning strategy by applying an adaptive submodular optimization scheme [8]. Motivated by this line of submodular optimization techniques, our active

learning method firstly explores and leverages structured information of histopathological images through imposing a partition matroid constraint on active learning.

2 Approach

2.1 Problem Definition

Given an unlabeled dataset $\mathcal{U} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, each data sample $\boldsymbol{x}_i \in \mathcal{U}$ carries a random label variable $y_i \in \mathcal{Y}$ ($\mathcal{Y} = \{1, -1\}$) in our binary classification task for which the positive label '1' implies 'benign' and the negative label '-1' implies 'actionable'. Assume that there exists a joint probability distribution $P(\boldsymbol{y}_\mathcal{U})$ of the labels of the samples in $\mathcal{U}$, where $\boldsymbol{y}_\mathcal{U} = [y_1, \cdots, y_n]^\top \in \mathcal{Y}^n$. Batch mode active learning selects a small subset of $\mathcal{U}$, queries their labels from experts, and then trains a classifier using the chosen labeled samples. To be specific to histopathological image analysis, batch mode active learning works as follows: whenever a batch of k unlabeled images $\mathcal{B} \subseteq \mathcal{U}$ ($|\mathcal{B}| = k$) are selected, their associated labels $\boldsymbol{y}_\mathcal{B} \in \mathcal{Y}^k$ are requested from the diagnosis of pathologists and acquired simultaneously; the obtained labels are used to select next batches of images iteratively until the needed classification (*i.e.*, predicting 'benign' or 'actionable') accuracy is achieved.

2.2 Adaptive Submodular Optimization

Our goal is to learn a classifier $h : \mathcal{U} \to \mathcal{Y}$ from a set $\mathcal{H}$ of finite hypotheses. We write $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\} \subseteq \mathcal{U} \times \mathcal{Y}$ to denote the set of observed sample-label pairs. We define $\mathcal{H}(\mathcal{S}) = \{h \in \mathcal{H} : y_i \equiv h(\boldsymbol{x}_i), \forall (\boldsymbol{x}_i, y_i) \in \mathcal{S}\}$ to denote the reduced hypothesis space consistent with the observed sample-label pairs in $\mathcal{S}$. We then define and aim to maximize the objective set function $f : 2^{\mathcal{U}\times\mathcal{Y}} \to \mathbb{R}$ as

$$f(\mathcal{S}) = |\mathcal{H}| - |\mathcal{H}(\mathcal{S})|, \tag{1}$$

where the operator $|\cdot|$ outputs the cardinality of an input set. In this paper, we study hyperplane hypotheses in the form of $h(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x})$ in which the sign function $\mathrm{sgn}(x)$ returns 1 if $x > 0$ and -1 otherwise. Intuitively, the function $f(\mathcal{S})$ measures the number of hypotheses eliminated by the observed labeled data in $\mathcal{S}$. As a matter of fact, f satisfies the following properties:

- $f(\varnothing) = 0$; (**Normalized**)
- for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$, $f(\mathcal{S}_1) \leq f(\mathcal{S}_2)$; (**Monotonic**)
- for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$ and $(\boldsymbol{x}, y) \in (\mathcal{U} \times \mathcal{Y}) \backslash \mathcal{S}_2$, we have $f(\mathcal{S}_2 \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S}_2) \leq f(\mathcal{S}_1 \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S}_1)$; (**Submodular**)
- for an unlabeled sample $\boldsymbol{x}$ and an observed data subset $\mathcal{S} \subseteq \mathcal{U} \times \mathcal{Y}$, define the conditional expected marginal gain of $\boldsymbol{x}$ with regard to $\mathcal{S}$ as

$$\Delta_f(\boldsymbol{x} \mid \mathcal{S}) = \sum_{y \in \mathcal{Y}} P(y_i = y \mid \mathcal{S})[f(\mathcal{S} \cup \{(\boldsymbol{x}, y)\}) - f(\mathcal{S})], \tag{2}$$

  and then the function f along with the distribution $P(\boldsymbol{y}_\mathcal{U})$ is called adaptive submodular if $\Delta_f(\boldsymbol{x} \mid \mathcal{S}_2) \leq \Delta_f(\boldsymbol{x} \mid \mathcal{S}_1)$ holds for any $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{U} \times \mathcal{Y}$ and $P(\mathcal{S}_2) > 0$. (**Adaptive Submodular** [8])

To work under the batch mode setting, the *BatchGreedy* algorithm [2] generalizes the conditional marginal benefit in Eq. (2) to allow for conditioning on a set of selected but not yet observed sample-label pairs within the current batch. *BatchGreedy* greedily selects the samples within each batch and assembles batches in a sequential manner. Specifically, *BatchGreedy* selects the i-th sample in the j-th batch as follows:

$$\boldsymbol{x}^* = arg\max_{\boldsymbol{x}\in\mathcal{U}} \Delta_f(\boldsymbol{x} \mid \{\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{i-1,j}\}, \mathcal{S}), \tag{3}$$

where $\mathcal{S}$ represents the observed labeled data from all previous $j-1$ batches, and $\{\boldsymbol{x}_{1,j}, \cdots, \boldsymbol{x}_{i-1,j}\}$ retains the selected $i-1$ samples whose labels are not observed yet within the current j-th batch. This algorithm is theoretically guaranteed to obtain an approximation to the optimal batch-mode active sampling strategy.

2.3 Modeling the Partition Matroid Constraint

Since images of the same patient are very likely to include large pathological information redundancy, we propose to explicitly enforce diversity within the selected images by imposing an additional partition matroid constraint on the original adaptive submodular function maximization problem in Eq. (3).

A partition matroid constraint is defined as follows: $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_q$ are a partitioning of the set $\mathcal{U}$ if $\mathcal{U} = \bigcup_{1\le i\le q} \mathcal{P}_i$ and $\mathcal{P}_1, \cdots, \mathcal{P}_q$ are disjoint with each other. We require the currently selected batch to include at most one sample from each subset $\mathcal{P}_i$.

More formally, our proposed constrained problem is defined as follows:

$$\begin{aligned} \mathcal{B}^* = \ & \arg\max_{\mathcal{B}\subseteq\mathcal{U}} \Delta_f(\mathcal{B} \mid \mathcal{S}) \\ & \text{subject to } |\mathcal{B}| = k, |\mathcal{B}\cap\mathcal{P}_i| \le 1, k \le q, \forall i \in \{1, ..., q\}, \end{aligned} \tag{4}$$

where $\mathcal{B}^*$ is the optimal k-cardinality batch selected from the current unlabeled dataset $\mathcal{U}$, $\mathcal{P}_1, \cdots, \mathcal{P}_q$ are q disjoint subsets partitioning $\mathcal{U}$, and $\mathcal{S}$ is the set composed of the previously observed labeled data. These disjoint subsets can be obtained through performing clustering according to the structured information of the annotated images.

Within each batch, the i-th sample of the j-th batch is selected as follows

$$\begin{aligned} \boldsymbol{x}^* = \ & \arg\max_{\boldsymbol{x}\in\mathcal{U}} \Delta_f(\boldsymbol{x} \mid \{\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{i-1,j}\}, \mathcal{S}) \\ & \text{subject to } cluster(\boldsymbol{x}) \neq cluster(\boldsymbol{x}_{k,j}), \forall k \in \{1, \cdots, i-1\}, \end{aligned} \tag{5}$$

where $cluster(\boldsymbol{x})$ is the index of the cluster that $\boldsymbol{x}$ belongs to.

For the sequential version of this problem, Golovin and Krause[7] have proven that the greedy method can achieve a $(\frac{1}{p+1})$-approximation to the optimum when maximizing f subject to p matroid constraints, which motivates us to generalize this result to the batch mode setting. We propose a practical batch mode active learning algorithm BGAL-PMC, as described in Algorithm 1. In what follows, we show that BGAL-PMC can well approximate the optimal batch selection strategy. Note that $\mathcal{H}$ is the hypothesis set, $\mathcal{H}(\mathcal{S})$ is the reduced hypothesis set which is consistent to the observation $\mathcal{S}$, and $|\mathcal{H}|$ is the size of the hypothesis set.

Algorithm 1. BGAL-PMC (Batch Greedy Active Learning with a Partition Matroid Constraint)

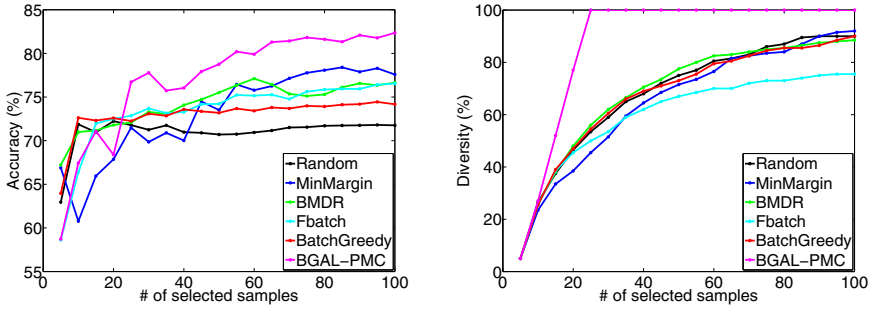
Input: a set of disjoint clusters $\mathcal{P}_1, \mathcal{P}_2, ...\mathcal{P}_q$, previously selected dataset $\mathcal{S}$ and their observed labels $\boldsymbol{y_S}$, unlabeled dataset $\mathcal{U}$, hypothesis set size N, and batch size k.
Ouput: the selected batch $\mathcal{B}$ and their labels $\boldsymbol{y_B}$.
Sample a hypothesis set $\mathcal{H} = \{h_1, h_2, ...h_N\}$ using $\boldsymbol{y_S}$;
initialize $\mathcal{B} = \emptyset$, $D = \emptyset$, and $\mathcal{T} = \emptyset$;
for $i = 1$ **to** k **do**
  for $j = 1$ **to** $|\mathcal{U}|$ **do**
    $score(\boldsymbol{x}_j) = \mathbb{E}_{y \in \{-1,1\}}[|\mathcal{H}(\{(\boldsymbol{x}, y) \mid \boldsymbol{x} \in \mathcal{B} \cup \{\boldsymbol{x}_j\}\})|]$
  end for
  while true **do**
    $\boldsymbol{x}^* = arg\min_{\boldsymbol{x} \in \mathcal{U} \setminus \{\mathcal{B} \cup \mathcal{T}\}} score(\boldsymbol{x})$
    $ind = cluster(\boldsymbol{x}^*)$
    if $ind \notin D$ **then**
      $\mathcal{B} = \mathcal{B} \cup \{\boldsymbol{x}^*\}$, $D = D \cup \{ind\}$
      break
    else
      $\mathcal{T} = \mathcal{T} \cup \{\boldsymbol{x}^*\}$
    end if
  end while
end for
query the labels $\boldsymbol{y_B}$ for $\mathcal{B}$.

Theorem 1. *Given a monotonic and submodular function f and a label distribution P such that (f, P) is adaptive submodular, when maximizing f subject to a partition matroid constraint, the expected cost of the BGAL-PMC algorithm is at most* $2(\ln(|\mathcal{H}| - 1) - 1)$ *times the expected cost of the optimal batch selection strategy.*

The proof of Theorem 1 is provided in the supplemental material. This theorem guarantees that BGAL-PMC needs at most $2(\ln(|\mathcal{H}| - 1) - 1)$ times more batches than those required by the optimal batch selection strategy. Note that directly searching for the optimal selection strategy takes exponential time. To sample a finite hypothesis set $\mathcal{H}$, we employ the hit-and-run sampler [11] to generate a set of linear separators, which has been used by [2] and proven effective for active learning problems.

3 Experiments

Experimental Settings: Our experiments are conducted on a large database of histopathological images from breast microscopic tissues [4,17]. This database contains more than two thousand images, gathered from around a hundred patients. Each image is labeled as benign category (usual ductal hyperplasia (UDH)) or actionable category (atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS)) by pathologists, which are development procedures from a normal terminal duct-lobular unit to an invasive cancer. Classifying these two categories is an important clinical problem since the therapy planning and management relies on the diagnosis of UDH



(a) Accuracy Comparison (b) Diversity Comparison

Fig. 1. (a) Learning curves of the proposed BGAL-PMC and other 5 methods on the breast microscopic tissues image dataset. X-axis is the number of selected images while Y-axis is the accuracy as the number of selected training images increases. BGAL-PMC (the pink curve) outperforms the other 5 methods significantly; (b) The diversity curves of all 6 methods. X-axis is the number of selected images while Y-axis is the diversity of the selected set as the number of selected images increases. Note that the diversity here is defined as the perentage of partitioning clusters being covered.

and ADH/DCIS. It is also very challenging due to the subtle differences between categories. High-dimensional (i.e., 10000) texture features are extracted from each image. We randomly split the dataset into 50% training to actively select candidate images and 50% testing to test the learned classifier. We also ensure that images for a particular patient are either in the training set or in the testing set. We randomly split 10 times and the average performance is reported.

Five active learning methods are compared, i.e., Random Selection, Min Margin [15], Fbatch [10], BMDR [16], and BatchGreedy [2]. Note that the Random Selection is equivalent to the passive learning setting. In our method, we partitioned the dataset into 20 disjoint subsets using both the structured information and image texture features by K-means. Since it's difficult and time-consuming to sample hyperplanes uniformly in high dimensional space, we follow [2] to reduce the dimension to 100 to sample the hypothesis set $\mathcal{H}$. For fair comparison, we use SVM classifier for all methods, with the same parameters tuned via five-fold cross validation. We set batch size at 5 throughout the experiments. Two positive images and two negative images are randomly selected for initialization. The size of the hypothesis set is set at 300, which is empirically large enough in our experiments. All experiments are conducted on a 2.80GHz i7 CPU with 16 cores and 16G RAM in Matlab implementation.

Results: Fig. 1(a) shows the classifier learning curves as selected samples increase. Not surprisingly, all five active learning methods perform better than random selection, which manifest the effectiveness of active learning. In particular, the proposed BGAL-PMC performs significantly better than all other four active learning methods. Min Margin method as a classical active learning baseline is the second-best in our experiments. Although Fbatch, BMDR and BatchGreedy perform well in the first 20 selected samples, the improvement of their accuracy is less substantial when more

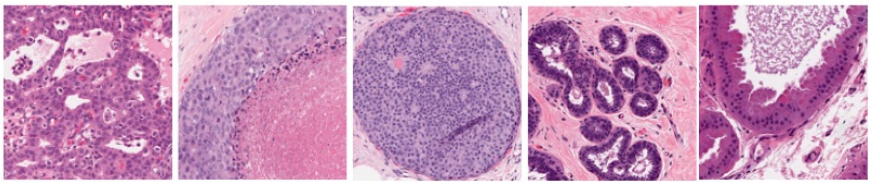


Fig. 2. One example batch of selected images using our proposed method. The first 3 are actionable, and the last 2 are benign. 5 images are selected from distinct clusters.

Table 1. Comparsion of the average time to select a single batch of images for 5 active learning algorithms (batch size=5)

Methods	MinMargin[15]	BMDR[16]	FBatch[10]	BatchGreedy[2]	BGAL-PMC
Time (seconds)	3.13	17.63	128.13	1.97	1.98

batches are selected. The reason is that all other methods do not take the information of clusters into consideration. Therefore, their selected images may include information redundancy, which downgrades their performances. On the other hand, trivially using cluster information cannot achieve the same accuracy either. We tested samping from randomly-chosen distinct clusters, as an alternative baseline. It achieved 77% accuracy when selecting 100 samples which is better than some baselines, but is still significantly worse than our proposed method. Leveraging image structured information may be a general paradigm to boost active learning performance, but our proposed matroid constraint is a more effective and theoretical sound method. With less than 5% data labeled, our method achieves 83% prediction accuracy. This accuracy is at least 6% higher than all compared methods. In fact, when 80% data is labeled, the prediction accuracy is 87%, which is merely 4% higher than our method but use much more labeled samples than us. Therefore, this scheme considerably reduces the label effort from pathologists, without significantly sacrificing the accuracy.

We further investigated the diversity of all methods, as shown in Fig. 1(b). The diversity here is defined as the coverage rate of the clusters. Since we enforce the partition matroid constraint explicitly, BGAL-PMC covered all the clusters in much fewer iterations than other methods. Fig. 2 is one selected batch using our proposed method, to show the diversity of our selections visually. We also compared the running time, as shown in Table 1. In our experiments, BatchGreedy and BGAL-PMC are much more scalable than other active learning algorithms. BatchGreedy is slightly faster than ours (1.97s vs. 1.98s), both of which are negligible in the practical use of active learning.

4 Conclusion

In this paper, we proposed a novel batch mode active learning approach which leverages the structured information of annotated histopathological images. We formulated the batch mode active learning problem as a submodular function maximization problem with a partition matroid constraint, which prompts us to design an efficient greedy

algorithm for approximate combinatorial optimization. We further provided a theoretic bound characterizing the quality of the solution achieved by our algorithm. We compared the proposed active learning approach against several state-of-the-art active learning methods on a large database of histopathological images, and demonstrated the superiority of our approach in performance. The spirit of our active learning method capitalizing on submodular optimization is generic, and can thus be applicable to other problems in medical image analysis. In the future, we will also explore more sophiscated ways to extract structured infomation.

References

1. Balcan, M.F., Hanneke, S., Vaughan, J.W.: The true sample complexity of active learning. Machine Learning 80(2-3), 111–139 (2010)
2. Chen, Y., Krause, A.: Near-optimal batch mode active learning and adaptive submodular optimization. In: Proc. ICML (2013)
3. Doyle, S., Agner, S., Madabhushi, A., Feldman, M., Tomaszewski, J.: Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In: Proc. ISBI (2008)
4. Dundar, M.M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O., Gurcan, M.N.: Computerized classification of intraductal breast lesions using histopathological images. IEEE Transactions on Biomedical Engineering 58(7), 1977–1984 (2011)
5. Fisher, M.L., Nemhauser, G.L., Wolsey, L.A.: An analysis of approximations for maximizing submodular set functions–ii. In: Polyhedral Combinatorics pp. 73–87 (1978)
6. Foran, D.J., Yang, L., et al.: Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. JAMIA 18(4), 403–415 (2011)
7. Golovin, D., Krause, A.: Adaptive submodular optimization under matroid constraints. arXiv preprint arXiv:1101.4450 (2011)
8. Golovin, D., Krause, A.: Adaptive submodularity: Theory and applications in active learning and stochastic optimization. JAIR 42(1), 427–486 (2011)
9. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering 2, 147–171 (2009)
10. Hoi, S.C., Jin, R., Zhu, J., Lyu, M.R.: Batch mode active learning and its application to medical image classification. In: Proc. ICML (2006)
11. Lovász, L.: Hit-and-run mixes fast. Mathematical Programming 86(3), 443–461 (1999)
12. Petushi, S., Garcia, F.U., Haber, M.M., Katsinis, C., Tozeren, A.: Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Medical Imaging 6(1), 14 (2006)
13. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
14. Settles, B.: Active learning literature survey. Technical Report, University of Wisconsin, Madison (2010)
15. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2002)
16. Wang, Z., Ye, J.: Querying discriminative and representative samples for batch mode active learning. In: Proc. KDD (2013)
17. Zhang, X., Liu, W., Zhang, S.: Mining histopathological images via hashing-based scalable image retrieval. In: ISBI (2014)