An Efficient Many-Class Active Learning Framework for Knowledge-Rich Domains

Weishi Shi and Qi Yu
Rochester Institute of Technology
Rochester, New York
email: {ws7586, qi.yu}@rit.edu

***Abstract*—The high cost for labeling data instances is a key bottleneck for training effective supervised learning models. This is especially the case in domains such as medicine and bioinformatics, where expert knowledge is required for understanding and extracting the underlying semantics of data. Active learning provides a means to reduce human labeling efforts by identifying the most informative data instances. In this paper, we propose a cost-effective active learning framework to further lessen human efforts, especially in knowledge-rich domains where a large number of classes may be subject to scrutiny during decision making. In particular, this framework employs a novel many-class sampling model, MC-S, for data sample selection. MC-S is further augmented with convex hull-based sampling to achieve faster convergence of active learning. Evaluation studies conducted over multiple real-world datasets with many classes demonstrate that the proposed framework significantly reduces the overall labeling efforts through fast convergence and early stop of active learning.**

***Index Terms*—active learning; data sampling; knowledge-rich domains**

I. Introduction

Obtaining labeled data in knowledge-rich domains (e.g., bioinformatics and medicine) is usually challenging and expensive. For example, labeling a data instance in the medical domains is equivalent to making a diagnosis based on a patient's medical condition on file. Such annotation process heavily relies on physicians' domain knowledge that is obtained through years of medical training. The difficulty of collecting training data in knowledge-rich domains stimulates the high drive for an efficient computational framework to reduce the overall data labeling costs.

Active learning (AL) provides an effective means to reduce human labeling efforts by selecting the most informative data instances. It has been successfully applied in various applications [1]–[3]. One fundamental question in active learning is how to choose the most informative data instances from a candidate pool. Ideally, these data instances should contribute most to the increase of model accuracy, and practically a classic active learner selects the instances which are the most confusing to the classifier. A Support Vector Machines (SVMs) classifier has been widely used in active learning as it provides a convenient way to choose confusing data instances from the unlabeled pool [4]–[6]. For a typical binary-class problem, the selected data instances are those closest to the current decision boundary. This simple strategy, as well as its variations, achieves high accuracy efficiently [4].

In knowledge-rich domains, such as dermatology and radiology, extracting semantics from data instances requires much expert knowledge. Due to the complexity of the body of knowledge, decision making in these domains may involve a large number of classes (the *many-class* problem). For this reason, directly applying active learning in knowledge-rich domains introduces additional challenges: (1) From the machine perspective: The classic active learning is not specially designed to select data samples that efficiently update as many decision boundaries. Due to the interplay of a large number of classes, the decision boundaries can be very complicated. This undoubtedly makes data sample selection more challenging, in spite of some existing approaches that can deal with multiple classes [7]. As a consequence, the classic active learning suffers from slow convergence. (2) From the expert perspective: Labeling a single data instance may become nontrivial, as a large number of classes are candidates, where each requires serious inspection of domain-specific details. This affects the performance of domain experts, because high mental workload causes fatigue [8].

To address the aforementioned challenges, the framework proposed in this paper benefits the application fields and contributes to the literature as follows:

- We develop a *Many-Class Sampling (MC-S)* model that prefers a data sample that is both confusing in terms of the predicted class label and uncertain over the remaining classes. By achieving a good balance between *confusion* and *uncertainty*, MC-S selects the data samples that are most effective to improve the decision boundaries of a large number of classes.
- We further define a *unified objective function* that allows choosing data samples with the potential to significantly change the current model. Data sampling is achieved by solving a convex optimization problem, which can be done efficiently. Meanwhile, by monitoring the model change, active learning can terminate early without being tested on a hold-out dataset, which further reduces the labeling efforts.

We conducted evaluation experiments over multiple real-world datasets from diverse domains with many classes. The experimental results demonstrate the effectiveness and efficiency of the proposed many-class active learning framework.

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00164

II. Related Work

We group the recent studies on active learning to reduce human data labeling efforts into two categories, and in each category we compare the existing studies with our proposed approach to highlight the key differences.

A. Candidate Data Instance Sampling

To address the multi-class problem, Culotta and McCallum extended the uncertainty sampling by first identifying the dominant class of each data instance and then selecting among all instances the one whose dominant class has the smallest probability [9]. However, this approach may be stuck with selecting instances whose class probabilities are evenly distributed. As an improvement, a marginal sampling strategy was developed to choose the data instance with minimal difference between its most and second most probable classes, a.k.a., Best-versus-Second Best (BvSB) model [7]. However, BvSB only considers the two most probable classes (the local pairwise class distribution), which ignores the probability distribution of other classes. This makes it less effective when a large number of classes are involved. Entropy-based sampling is shown to be effective to quantify uncertainty for multi-class active learning in multiple studies [10], [11]. However, this sampling approach can be unstable when no precise entropy information can be provided due to lack of instances in each class at the beginning of active learning. This could be even worse in the case where there are a large number of classes. Probabilistic models provide an alternative way of considering all potential classes. Kottke et al. proposed a multi-class probabilistic active learning model (McPAL), which computes the expectation of the classification error within a neighbourhood of a candidate as a sampling score [12]. A closed-form solution is developed for efficient expectation computation. However, our empirical study shows that the computational cost still increases significantly with the number of classes, making it infeasible to handle many-class problems in practice.

The proposed MC-S sampling aims to address the above issues by choosing data samples that are effective to improve the decision boundaries of a large number of classes. A convex hull-based objective function is also developed to guide the sampling process so that the entire sample space can be efficiently explored to ensure fast convergence.

B. Termination Criterion

A common way to determine a termination criterion of active learning is to estimate model confidence on a holdout validation dataset [13], [14]. This approach suffers from late termination and hence requires more labeling efforts [15]. In addition, extra human efforts are needed to label the validation set [16], making it less attractive for knowledge-rich domains. As a substitute, sample diversity-based approaches do not depend on a labeled validation set [17], [18]. However, they instead require solving a convex optimization problem on the entire unlabeled data pool to maximize sample diversity, which is inefficient in case of a large-scale pool. Termination can also be indicated by some model properties [4], [6]. However, these approaches tend to stop late as they rely on high-level statistical summary of the model rather than on the localized learning behaviors that can be obtained during active learning.

In this paper, we propose automatic termination criteria based on data sample skipping and convex approximation error of selected samples using a small subset of labeled data, respectively. They both achieve early termination in multiple evaluation studies.

III. Knowledge-Rich Datasets

We provide a detailed description of two datasets collected from a knowledge-rich domain. These are two transcribed speech corpora that are collections of dermatology diagnostic narration documents. **Derm 1** is collected by instructing 16 participating physicians to describe each image content toward a diagnosis. The 50 dermatology images (50 diagnoses) form a total of 50 classes. **Derm 2** is collected with 29 physicians and 30 images (classes). For each class, there are 29 data instances, each from one physician. The narration documents are of drastically different lengths due to the narration styles used by different physicians. A large portion of each narration document is formed by specialized medical terms whose meanings can only be interpreted by experts.

The dermatology corpora are an ideal testbed for our active learning framework for these reasons: (1) Dermatology is a medical specialty that highly depends on specialized skills obtained through years of training. Recruiting the appropriate experts to label dermatology data is both challenging and expensive. Reducing document labeling costs in this domain showcases a tight connection of the proposed framework to the real-world clinical settings. (2) Both datasets contain a relatively large number of potential classes (diagnoses) to choose from for labeling. This highlights the major advantages of our learning framework against its counterparts. (3) Successful application to such highly specialized and challenging datasets helps demonstrate the applicability of our model to other datasets and domains where similar challenges arise.

IV. The Active Learning Framework

In this section, we first describe a basic model for multi-class SVM active learning. We proceed to describe the proposed many-class sampling (MC-S) model and highlight its key difference and advantage. We then present a key extension of MC-S, which leverages a convex hull-based unified objective function for data sampling (MC-CH).

A. Multi-Class SVM Active Learning

An SVM active leaner uses the distance to the separating hyperplane as a way to choose the most informative data sample to label. However, this criterion is not directly applicable to multi-class problems given the interplay of multiple hyperplanes. A principled approach is to compute the posterior probabilities of all the classes and use them to guide data sample selection. While most probabilistic models provide a natural way to generate the posterior probabilities, we choose

an SVM classifier since some important properties of the support vectors (i.e., sparsity and closeness to the decision boundary) can further benefit the active learning process. In particular, Platt scaling [19] and pairwise coupling [20] are used to convert the decision function of an SVM to posterior probabilities of classes.

B. Many-Class Sampling (MC-S)

Consider a pool of M unlabeled data samples: $X \in \mathbb{R}^{M\times N}$, where N denotes the number of features. The probabilistic output of an SVM classifier can be denoted by a matrix $C \in \mathbb{R}^{M\times K}$, where K is the number of classes and $C_{i,j} = p(C_j|\mathbf{x}_i)$ and $\sum_{j=1}^{K} C_{i,j} = 1, \forall i \in [1, M]$. The predicted label of $\mathbf{x}_i$ is given by

$$\hat{y}_i = \arg\max_j p(C_j|\mathbf{x}_i) = \arg\max_j C_{i,j} \tag{1}$$

We refer to a prediction as a *non-confusing* one if $C_{i,\hat{y}_i} \gg \max_{j\neq\hat{y}_i} C_{i,j}$, which implies that $\mathbf{x}_i$ is located on the far positive side of the class $\hat{y}_i$ and far negative side of other classes. Since a non-confusing sample is far away from the decision surface of the SVM, adding it into the training set will not typically improve the current classifier. Based on this, Joshi et. al [7] developed a sampling mechanism, referred to as Best-versus-Second Best (BvSB), which chooses the most *confusing* data sample using the following rule:

$$\arg\min_i \ (C_{i,\hat{y}_i} - \max_{j\neq\hat{y}_i} C_{i,j}) \tag{2}$$

A sample is selected when the posterior probability of its predicted class has the smallest difference from that of its most competitive class. In essence, BvSB chooses the data sample that is most confusing for label assignment. For $K = 2$, BvSB reduces to the binary SVM active learner [7]. Generalized to the multi-class case, BvSB ensures the decision boundary of the two most probable classes to be effectively updated by sampling the most confusing data sample. However, its impact on other classes is not guaranteed and can be minor. Consequently, this requires more data instances to update the model for all classes, which is less effective, especially in case of a large number of classes.

1) Sampling Rule of MC-S: The proposed MC-S model addresses this issue by considering both local pairwise class distribution and the global class distribution of all classes. More specifically, we can use

$$p(\mathcal{R}_i) = 1 - C_{i,\hat{y}_i} - \max_{j\neq\hat{y}_i} C_{i,j} \tag{3}$$

$$\mathcal{R}_i = \{k \in 1, .., K | k \neq \hat{y}_i, k \neq \arg\max_{j\neq\hat{y}_i} C_{i,j}\} \tag{4}$$

to denote the chance of updating all the remaining classes. Note that $\max p(\mathcal{R}_i) = (K-2)/K$ is obtained when $H(\tilde{C}_i)$ is maximized where $H(\tilde{C}_i)$ is the entropy of random variable $\tilde{C}_i$ denoting the predicted posterior probabilities of $\mathbf{x}_i$. Therefore, MC-S uses the following rule for data sampling:

$$\arg\min_i \ F_{MC-S} = (C_{i,\hat{y}_i} - \max_{j\neq\hat{y}_i} C_{i,j}) + \lambda \sum_{j=1}^{K} C_{i,j} \log C_{i,j} \tag{5}$$

where the second term is the negative entropy ($-H(\tilde{C}_i)$) as we try to minimize the entire objective function. The sampling rule in Eq. (5) aims to choose a data sample that is both confusing in the predicted label and uncertain over the remaining classes. The first term ensures that the decision boundary around the predicted class will be significantly updated while the second term allows a decent chance for other classes' decision boundaries to be updated, making it more effective in a many-class situation.

2) Dynamic Update of λ: In the initial phase of active learning, since the model is not well trained yet, the entropy term may be estimated very inaccurately. As active learning continues and the model accuracy keeps improving, the entropy term should play a more important role as it helps choose data samples that are uncertain over a large number of classes (hence improve their decision boundaries if being labeled). We propose to dynamically adapt λ according to the progression of active learning. The rationale is that the model accuracy is expected to be higher when more data instances are labeled. More specifically, let λ_0 denote the initial value of λ (which is set to 0.7 in our experiments to give more weight to the first term in Eq. (5) in the initial phase of active learning) and n denote the iteration number of active learning. The update rule is given by

$$\lambda = \lambda_0 + \left\lfloor \frac{n}{K} \right\rfloor \times r \tag{6}$$

where r is the increasing rate, which is set to 0.05 in our experiments but other values in a similar range work equally well. In essence, this update rule increases the weight of the entropy term by 0.1 after every $2K$ samples are labeled.

C. Convex Hull-based Unified Sampling

The MC-S model relies on the current decision boundaries for sampling, making it sensitive to the initialization of the active learner. It also tends to choose data samples that are close to the current decision boundaries and hence is less effective to explore the entire data sample space. These may limit the convergence speed of MC-S.

These issues can be addressed if we avoid labeling the data samples that are less effective to update the current decision boundaries. We refer to the data samples selected by the active learner that do not bring significant changes to the current decision boundaries as *non-sensitive* data samples. These samples can be identified by checking whether their contribution to the decision boundaries can be approximated by existing support vectors. The reason of using support vectors to approximate the new data sample is two-fold: (1) they are close to the decision boundaries and comparing with them allows us to assess how much the data sample may change the decision boundaries, and (2) they are sparse, which guarantees good efficiency.

A straightforward way to implement the idea above is to check whether we can find a support vector from the predicted most probable class that is close enough to a selected data sample. If so, we skip this sample from labeling it. We refer this strategy as Nearest Neighbor-based Sample Skipping

(MC-NN). However, there are two limitations with such an approach: (1) Its effectiveness for identifying non-sensitive data samples may be limited when the support vectors are very sparse. (2) Some useful data samples may be wrongly skipped due to the inaccurate prediction of their labels. Since these data samples are permanently skipped and never got labeled, MC-NN may lead to a much lower model accuracy at the end of the active learning process, which is not desired.

We present a convex hull-based unified sampling function (MC-CH), which addresses the above issues while ensuring a fast convergence of active learning. The MC-CH is motivated by the following theorem.

Theorem 4.1: Adding a data sample that falls into the convex hull of the support vectors from the same class does not change the decision boundary of an SVM. ∎

The theorem can be proved by verifying the optimal separating hyperplane of the SVM and the KKT condition remain unchanged after adding a data sample described above (detailed proof is skipped due to the lack of space). Instead of using individual support vectors, MC-CH leverages all the support vectors in the predicted class to increase the chance of skipping non-sensitive data samples. Different from MC-NN that focuses on a local neighborhood, MC-CH considers the overall geometric structure of the decision boundaries. It uses the convex hull of the support vectors to approximate the decision boundary of the predicted class.

1) Avoiding Wrong Skips: MC-CH addresses the wrong skipping issue of MC-NN by penalizing the non-sensitive data samples instead of skipping them. The penalty may only slightly postpone the labeling of useful data samples when the model becomes more accurate. Therefore, the model will still benefit from labeling those data samples, which ensures convergence to a high accuracy (see Fig. 2 for the result).

The key idea of penalty-based sampling is to add a penalty term to the many class sampling rule in Eq. (5), where the penalty term is proportional to the distance of a data sample to the support vectors from the same predicted class. Given a candidate data sample $\mathbf{x}'$ and its predicted label, its distance to the convex hull of all the support vectors in the same predicted class can be measured through the residual error by using the convex combination of these support vectors to approximate $\mathbf{x}'$. More formally, let $S = (\mathbf{s}_1, ..., \mathbf{s}_k)$ denote the support vectors of interest and $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^\top$ denote the combination coefficients. The approximation error function of $\mathbf{x}'$ is given by $e(\mathbf{x}'; \boldsymbol{\theta}) = ||S^\top \boldsymbol{\theta} - \mathbf{x}'||^2$. To determine the minimum approximation error, denoted by $\hat{e}(\mathbf{x}') = \min_{\boldsymbol{\theta}} e(\mathbf{x}'; \boldsymbol{\theta})$, we solve the following quadratic (and convex) problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} e(\mathbf{x}'; \boldsymbol{\theta})$$
$$\text{subject to} \quad \theta_i \geq 0, (i = 1, ..., k) \quad \sum_{i=1}^{k} \theta_i = 1 \tag{7}$$

A small approximation error $\hat{e}(\mathbf{x}')$ indicates that $\mathbf{x}'$ stays close to the convex hull of its predicted class. This implies that $\mathbf{x}'$ will not significantly change the current decision boundary and hence should be penalized to reduce its chance of being sampled. Conversely, a large error means that $\mathbf{x}'$ is far from all the support vectors in the predicted class. In this case, $\mathbf{x}'$ is expected to bring a significant enough change to the decision boundary to achieve fast convergence, which should give it a better chance to be sampled. Since MC-S samples data by minimizing F_{MC-S}, we propose to use $-\hat{e}(\mathbf{x})$ as the penalty term, which leads to the unified sampling rule:

$$\begin{aligned} \mathbf{x}' &= \arg\min_{\mathbf{x}} F_{MC-CH}(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} F_{MC-S}(\mathbf{x}) - \gamma \hat{e}(\mathbf{x}) \end{aligned} \tag{8}$$

where $\mathbf{x}$ denotes each data sample in the current unlabeled pool. The first term in Eq. (8) tends to choose data samples that are effective to refine multiple decision boundaries of the current model while the second term sets the further preference to the samples with a potential to dramatically change the model. In our experiments, we use the same dynamic update rule of λ to update γ.

D. Automatic Termination Detection

Both MC-NN and MC-CH allow automatic termination of active learning without relying on a hold-out labeled dataset, making them more attractive for knowledge-rich domains. Specifically, MC-NN skips all the non-sensitive data samples. Active learning is terminated when all the unlabeled samples have been visited by MC-NN. Since a large number of data samples are skipped without being labeled, early termination can usually be achieved (see Fig. 2 for the result).

Instead of skipping data samples, MC-CH computes the approximation error of data samples using support vectors of the predicted class. At the convergence of active learning, the decision boundaries become stable. This implies that the new data samples stay close to the convex hulls of their respective classes and hence can be well approximated by the convex combination of their support vectors. Therefore, we observe a significant drop of the approximation error, which serves as an important indicator to terminate the active learning process as early as possible. Our experimental results confirm this (see Fig 2) and provide empirical evidence that MC-CH automatically detects the stopping condition of active learning without relying on a labeled holdout dataset.

V. Experiments

We have conducted extensive experiments to evaluate the proposed many-class active learning framework. The evaluation covers the following major aspects: many-class sampling performance and effectiveness of convex hull-based unified sampling function.

A. Datasets and Settings

Besides the two dermatology corpora as described in Section III, the experiments also include four additional datasets with a decent number of classes. These datasets are collected from diverse domains and evaluation over them help demonstrate the general applicability of the proposed active learning framework. Table I summarizes the major characteristics of all datasets. Below is a brief description of additional datasets:

Dataset	#Inst	#Attr	#Classes	Class Distr.	Domain
Derm 1	800	1391	50	Even	Medical
Derm 2	868	1554	30	Even	Medical
Penstroke	1144	500	26	Even	Image
Yeast	1484	8	10	Skewed	Biology
Auto-drive	58509	48	11	Even	Auto
Reuters	10788	5227	75	Skewed	News

TABLE I: Description of Datasets

- **Penstroke** is comprised of images of hand-written English characters by people with distinct writing styles.
- **Yeast** is a biological dataset that consists of localization sites of proteins in bacteria.
- **Auto-drive** aims to predict abnormal conditions of automobiles without implementing additional sensors.
- **Reuters** is from the text domain that consists of a large collection of Reuters news reports.

To best reflect the high labeling cost for knowledge-rich domains, we use very limited labeled samples to initialize the active learning process. For relatively small datasets with evenly distributed classes, including the two dermatology and Penstroke datasets, one data sample per class is used. For Auto-drive, we use 20 labeled samples per class. For the two datasets with unevenly distributed classes, including Yeast and Reuters, we select 1% and 2% data samples from each class, respectively, according to the sizes of the datasets. All the labeled samples are randomly selected. The experiments are conducted three times with the averaged performance reported.

B. Active Learning Models for Comparison

We compare the proposed MC-CH active learning method with three competitive active learning models that can be applied to multi-class problems.

- **Best-vs-Second-Best (BvSB)** sampling method selects a data instance that minimizes the posterior difference between its most and second most probable classes [7].
- **Multi-class Probabilistic Active Learning (McPAL)** determines the sampling score of a data instance using a density weighted performance gain [12].
- **Entropy-based sampling method (Entr)** uses the Shannon entropy of the predicted class distribution of each candidate as the sampling score [21].

For McPAL, we use Radial Basis Function kernel (RBF) to compute the neighbor frequency vector. The length scale (δ) of the kernel function and the number of hypothetically considered labels (m) are set to 0.7 and 2, respectively, as suggested by the original paper. In contrast to other models, the sampling behavior of McPAL is independent from the choice of the classifier. Therefore, we use SVM rather than a Parzen window classifier or probabilistic KNN in the original paper to achieve a fair comparison with other models. Finally, random sampling (**Random**) is used as the comparison baseline.

C. Sampling Performance Comparison

Fig. 1 shows the comparison result over six datasets. An effective active learning algorithm is characterized by its fast converging rate, i.e., using less labeled samples to achieve a high model accuracy. MC-CH outperforms all its competitors

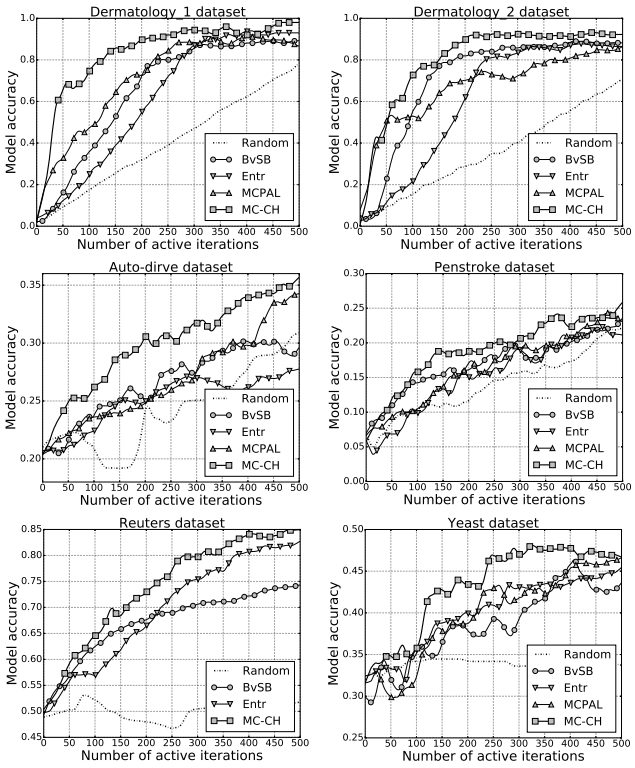


Fig. 1: Sampling Performance Comparison

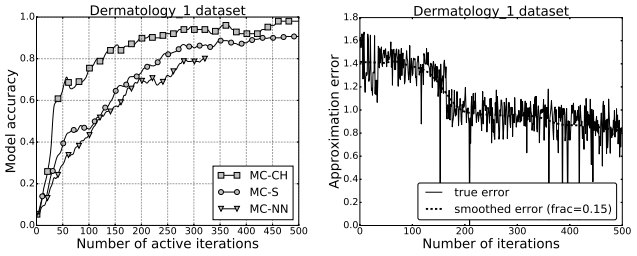
Dataset	MC-CH	[$\hat{e}(x)$ computed]	BvSB	Entr	MCPAL
Derm 1	0.55	[18.0%]	0.22	0.23	223.35
Derm 2	0.40	[8.2%]	0.12	0.13	127.24
Penstroke	0.32	[16.2%]	0.07	0.08	94.59
Yeast	0.85	[30.3%]	0.02	0.04	16.73
Auto-drive	0.62	[21.5%]	0.53	0.61	15.37
Reuters	19.35	[24.3%]	2.52	2.86	NA

TABLE II: Sampling Time Comparision

especially in the early and middle stages (before 250 iterations) of active learning. The performance advantage of MC-CH is due its two major contributors: (i) effectiveness of many-class sampling (MC-S) and (ii) the convex hull based unified function to best balance exploitation and exploration of the sampling space. The effects of these two contributors will be further investigated in the following subsection.

Besides the converging rate, we also report the sampling time of each active learning model in Table II. Compared with BvSB and Entr, the additional computation of MC-CH comes from the convex approximation error $\hat{e}(\mathbf{x})$ in (8). This computation can be further reduced as we use a lookup table to store and reuse $\hat{e}(\mathbf{x})$ as long as the predicted class and its support vectors are not changed for sample $\mathbf{x}$ in the candidate pool. Table II confirms the low percentage of data samples whose approximation errors need to be recomputed on average in each sampling iteration. Overall, all three models, including MC-CH, BvSB, and Entr, perform efficiently by using less than one sec for five out of six testing datasets. The relatively slow performance on Reuters is due to the large number of classes and the long news reports to be processed. With further increase of number of classes and candidate pool sizes, the sampling procedure can be expedited by testing

multiple samples simultaneously in parallel. While the number of classes plays an important role in the sampling time of each model, the sampling efficiency of McPAL decreases much more significantly due to the high cost of expectation computation. As a result, the sampling time of Reuters using McPAL is not reported due to the extremely slow process.



(a) MC-CH Vs. MC-NN, MC-S (b) Early Termination

Fig. 2: Effectiveness of MC-CH

D. Effectiveness of Convex Hull-based Unified Function

We investigate the effectiveness of MC-CH by further comparing it with MC-NN and MC-S in Figure 2a, where Derm 1 is used for illustration. For MC-NN, we use cosine similarity with a threshold of 0.2 to determine nearest neighbors. This threshold is set due to the sparsity of the narration documents and to achieve a good balance between skipping and model accuracy. A smaller similarity leads to more skipping and a lower model accuracy. The skipping rate also increases along with the active learning process. When the model approaches convergence, most data samples are skipped, which will eventually terminate active learning. As can be seen, MC-NN achieves automatic early termination after around 300 iterations. However, the low model accuracy of MC-NN confirms that it skips some very useful data samples. In contrast, MC-CH effectively addresses this issue by using a penalty term. It converges to the highest model accuracy but more efficiently than MC-S. Fig. 2b shows the early termination result by tracking the convex approximation error changes over active learning. To see a clear trend, we adopt Locally Weighted Scatter-plot Smoothing (LOESS) with fraction as 0.15 (i.e., 15% the data is used when estimating each y-value) to generate a smoothed curve. It can be seen that there is a significant drop of the smoothed error at around 200-th iteration and then the error becomes relatively stable. If we wait for another 50 iterations to make sure the error has been stabilized, we can stop at the 250-th iteration. Fig. 2a shows that the model is very close to its highest accuracy at this point. In fact, the remaining 250 samples, if being labeled, can only improve the model accuracy by 1%.

VI. Conclusions

In this paper, we present a novel active learning framework for the knowledge-rich domains to tackle the many-class problem which appears during data labeling. To update multiple classes' decision boundaries effectively and efficiently, this framework leverages an MC-S model to select data samples. MC-S is augmented with convex hull-based sampling (MC-CH) to achieve faster convergence of active learning. Automatic early termination of active learning is achieved by monitoring the change of convex approximation error, which avoids additional labeled data for validation. Extensive experiments conducted over multiple real-world many-class datasets clearly justify the effectiveness of the proposed active learning framework.

Acknowledgement

We would like to thank Anne Haake for providing the dermatology data. Both Anne and Xuan Guo provided helpful comments that improve the quality of the paper. This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

[1] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *KDD*, 2009, pp. 917–926.

[2] O. Mac Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," in *CVPR*, 2014, pp. 564–571.

[3] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *TIST*, vol. 2, no. 2, p. 10, 2011.

[4] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, 2000, pp. 839–846.

[5] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *JMLR*, vol. 2, no. Nov, pp. 45–66, 2001.

[6] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *CIKM*, 2007, pp. 127–136.

[7] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *CVPR*, 2009, pp. 2372–2379.

[8] W. Karwowski, *International Encyclopedia of Ergonomics and Human Factors*. CRC Press, 2001, vol. 3.

[9] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *AAAI*, vol. 5, 2005, pp. 746–51.

[10] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "Entropy-based active learning with support vector machines for content-based image retrieval," in *ICME*, vol. 1, 2004, pp. 85–88.

[11] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *CVPR Workshop*. IEEE, 2008, pp. 1–8.

[12] D. Kottke, G. Krempl, D. Lang, J. Teschner, and M. Spiliopoulou, "Multi-class probabilistic active learning." in *ECAI*, 2016, pp. 586–594.

[13] M. Li and I. K. Sethi, "Confidence-based active learning," *TPAMI*, vol. 28, no. 8, pp. 1251–1261, 2006.

[14] F. Laws and H. Schätze, "Stopping criteria for active learning of named entity recognition," in *ICCL*, 2008, pp. 465–472.

[15] M. Bloodgood and K. Vijay-Shanker, "A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping," in *CoNLL*, 2009, pp. 39–47.

[16] A. Vlachos, "A stopping criterion for active learning," *Computer Speech & Language*, vol. 22, no. 3, pp. 295–312, 2008.

[17] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *IJCV*, vol. 113, no. 2, pp. 113–127, 2015.

[18] E. Elhamifar, G. Sapiro, A. Yang, and S. Shankar Sasrty, "A convex optimization framework for active learning," in *ICCV*, 2013, pp. 209–216.

[19] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *ICML*, 2005, pp. 625–632.

[20] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *JMLR*, vol. 5, no. Aug, pp. 975–1005, 2004.

[21] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos, "Scalable active learning for multiclass image classification," *TPAMI*, vol. 34, no. 11, pp. 2259–2273, 2012.