Group-Based Active Learning of Classification Models

Zhipeng Luo and Milos Hauskrecht
Department of Computer Science
University of Pittsburgh, PA
zpluo, milos@cs.pitt.edu

Abstract

Learning of classification models from real-world data often requires additional human expert effort to annotate the data. However, this process can be rather costly and finding ways of reducing the human annotation effort is critical for this task. The objective of this paper is to develop and study new ways of providing human feedback for efficient learning of classification models by labeling groups of examples. Briefly, unlike traditional active learning methods that seek feedback on individual examples, we develop a new group-based active learning framework that solicits label information on groups of multiple examples. In order to describe groups in a user-friendly way, conjunctive patterns are used to compactly represent groups. Our empirical study on 12 UCI data sets demonstrates the advantages and superiority of our approach over both classic instance-based active learning work, as well as existing group-based active-learning methods.

1 Introduction

Learning of classification models from real-world data often requires additional human annotation effort that assigns class labels to data examples. Unfortunately, this process can be very costly and due to various budget and effort constraints, only a limited number of examples may be feasibly labeled, despite the fact that the unlabeled examples are abundant.

The most popular approach applied to build classification models from a limited human supervision effort is active learning. It aims to build a high-quality classifier by controlling what information is sought next. Active learning has gained popularity in many fields such as computer vision (Salmani and Sridharan 2014; Settles, Craven, and Ray 2008), natural language processing (Druck, Settles, and McCallum 2009; Small et al. 2011) and bio-medical data mining (Hoi et al. 2006; Haque et al. 2013; Valizadegan, Nguyen, and Hauskrecht 2013; XUE and Hauskrecht 2017).

Traditional active learning approaches assume that human feedback is instance-based. However in practice instance-based active learning may lead to imperfections and suboptimal performance. First, the instance-based active learning may be affected by the *sampling bias* problem (Dasgupta and Hsu 2008), in which the examples labeled throughout the active learning process are not good representatives of the overall distribution of examples and their labels. Second, instance-based approaches fail to fully leverage the diversity of existing human knowledge and to incorporate it in the model learning process; Finally, when instances are very complex high-dimensional data objects (such as, electronic health records) the review of each individual object (patient case) and its assessment may take a long time which rapidly increases the annotation cost.

In this work we explore a group-based active learning strategy to alleviate the above issues. We seek human label feedback on groups of instances. The strategy expects the user to assess the probability of one of the class labels in the subpopulation defined by the group. The main benefit of this approach, as opposed to instance-based learning, is that one may obtain information related to the class label on many different instances in just one query. Another benefit is that if data objects are very complex, the groups of instances may be defined more compactly by abstracting away many details of each individual case and hence easier and more efficient for human to review and assess. For example, when defining a concept of hemodynamic stability of a patient it is much easier for a clinician to consider patient subpopulations based only on heart rate and/or blood pressure ranges than a detailed clinical picture and intricacies of every possible patient observed in the past.

Based on this new query type, our active learning approach resembles a decision tree construction process where groups (rather than instances) are labeled in the top down fashion. We propose a splitting criterion called expected information gain, which is based on the group size and its estimated impurity, to refine the tree. We also show how the tree with softly labeled groups can be easily converted to a general instance-based classifier.

As an illustration example, suppose we want to build a classifier for predicting hospital admissions for the patients encountered in the Emergency Room (ER) based on the initial set of measurements such as heart rate, blood pressure, temperature. Initially, an ER clinician may estimate the chance of the admission to be 30% for the entire ER population, but this estimate may go up significantly to say 65% for the subpopulation with a high heart rate, and say 85% for a subpopulation of patients with both a high heart rate and a low blood pressure (that are the signs of significant blood loss). Our approach aims to take advantage of such subpopulations and their soft assessments to learn a classifier.

2 Related Work

Instance-Based Active Learning Enormous progress has been made on instance-based query strategies in recent years. Some are based on the uncertainty of instances (Lewis and Catlett 1994); Some rely on informativeness (Bodó, Minier, and Csató 2011) or representativeness (Nguyen and Smeulders 2004) of instances. Another body of instance-based work selects instances that could reduce the version space most, like query-by-disagreement and query-by-committee (Settles 2012). As mentioned in introduction section, however, instance-based active learning framework could suffer from the sampling bias problem (Dasgupta and Hsu 2008) and it is unfriendly to high-dimensional queries. Moreover, human supervision and feedback could be much more diverse than merely providing instance labels.

Alternative Query Types For all of the above reasons, alternative query types have been proposed and developed. For example, (Druck, Settles, and McCallum 2009) solicits information on NLP features rather than words (instances). Multiple Instance Active Learning, MIAL (Settles, Craven, and Ray 2008; Salmani and Sridharan 2014), uses two types of queries: bag-level and instance-level queries. MIAL follows the basic assumptions of Multiple Instance Learning framework (Amores 2013): a bag is a meaningful unity of instances, bag labels are asymmetric, and the goal is to train a bag-level classifier. However in our work, a group can represent an arbitrary set of instances, and soft labels (e.g 0.85 positive) are provided only on group-level. Also our goal is to learn an instance-level classifier as we are able to reinterpret group labels as instance labels.

Group-Based Active Learning There are two group active learning solutions that are most related our work: AGQ+ (Du and Ling 2010) and RIQY (Rashidi and Cook 2011). Although we have identical query types as they do, our framework is fundamentally different. Specifically, AGQ+ and RIQY extend traditional uncertainty sampling from single instance to multiple ones. In our framework, groups are formed hierarchically in a top-down manner. Moreover, we notice AGQ+ and RIQY have some limitations. Firstly, similar instances that are aggregated to form groups may not be actually 'similar' in the long run when the size of the unlabeled data shrinks, resulting in sparsity and 'holes' in the pool; Secondly, their data are only labeled once, which means labels can be permanently inaccurate. In our work instances can be labeled multiple times depending on the groups they reside in. Lastly, as criticized by the authors of RIQY, AGQ+ synthesizes numerous data instances for training. This step may not only bias the underlying distribution of data, it can also generate unrealistic data. For this reason, we only consider RIQY in our experimental evaluation section.

Other Active Learning Work Finally, it is also worth differentiating our work from other two active learning directions. The first one is active learning with clustering. In this line of work, (Nguyen and Smeulders 2004; Bodó, Minier, and Csató 2011) use clustering to measure the informativeness of instances; and (Dasgupta and Hsu 2008; Urner, Wulff, and Ben-David 2013) use hierarchical cluster structure to order and select instances. The second direction is batch-mode active learning, BMAL (Hoi et al. 2006; Chakraborty, Balasubramanian, and Panchanathan 2015; Guo and Schuurmans 2008). The essential difference of the work above from our work is that labels are still provided on instance level and therefore labeling cost is calculated over all examples. In our group-based framework, one query and one label is associated with the whole group.

3 Methodology

Our objective is to build a binary classification model to predict instance labels. We assume a pool of unlabeled data instances $\mathcal{U}$ is available at the beginning of the learning process.

Our active learning algorithm builds a decision tree that recursively partitions the pool $\mathcal{U}$ of examples and assigns a soft label to nodes (subpopulations or groups) represented in the tree. At the time when a given learning budget runs out, a classification model can be built from the subpopulations (groups) and their soft-labels that are the leaves of the tree. The key assumption of our work is that by gradually refining the groups in the tree (by increasing their number and reducing their size), the classification model learned from these groups can be improved.

3.1 Definitions and Notations

Given a pool of unlabeled training data, represented as a real number matrix $\mathcal{U}_{n\times m}$ which is composed of n m-dimensional instances, our goal is to learn a mapping $f : \mathbb{R}^m \mapsto \{-1,+1\}$. Instance $\mathbf{x}_i$ is the i-th row in $\mathcal{U}$, $\mathbf{x}_{id_j}$ denotes the value of $\mathbf{x}_i$ on dimension d_j.

Next we define the representation of a group in our framework. (group, tree node, hypercube and population may be used as inter-changeable concepts). A group G consists of three components: $G = (R_G,\ D_G,\ L_G)$, where R_G is a description of the group, $D_G \subset \mathcal{U}$ is a set of data points that belong to the group, and L_G is the soft label assigned.

The group description R_G is formed by a conjunctive pattern combining the attributes of the feature space $\mathcal{F} = \{d_1, ..., d_m\}$ and its values. For real-valued attributes individual patterns can be formed by inequalities, such as $2 < d_j < 12$. Inequalities are permitted also for ordinal category values. Patterns for categorical attributes are represented by equalities only. Then conjunctive patterns, such as, $(2 < d_j < 12 \wedge d_{j'} = BLUE)$ are used to describe groups. The main reason for choosing this description of the groups is to simplify the query interactions with humans. As AGQ+ and RIQY do, we measure the complexity of each group description by using feature reduction rate, which is defined as:

$$fr(R_G) = 1 - \frac{\#(used\ attributes\ in\ G)}{\#(All\ attributes)} \quad (1)$$

The data D_G belonging to the group G is a set of unlabeled data points that satisfy the group description R_G, that is, data points that fall into the region of the data space satisfying R_G. The label L_G in our framework is a soft label assigned to the group. If we assume the distribution of positive instances in the group G is binomial, then the soft label $L_G = \mu \in [0, 1]$ corresponds to the probability parameter of a binomial distribution in this group. In practice people may not be good at assigning very accurate μ, but in our experiments we show that a coarse level of precision is enough.

3.2 Active Group Learning Algorithm

In the following we present our algorithm for building the tree (hierarchy) of soft labeled groups from $\mathcal{U}$.

Our algorithm starts by forming the root group which is an unbound hypercube that covers the entire data space. Next we solicit its soft label, i.e. the prior of class distributions, from a human reviewer. The process continues by recursively splitting the leaf nodes of the current tree, hence refining the description of the examples represented by each tree node. Our algorithm iteratively seeks the leaf group with the greatest improvement potential and splits it into two disjoint groups. The sub-groups generated by the split are assessed by the reviewers and soft labels are assigned to the new groups. At the end, the algorithm will output a hierarchy of subpopulations and their associated soft labels.

In contrast to a typical decision tree learning process which is passive and instance-based, the exact implementation of our group-based active learning algorithm requires us to answer the following questions: 1) Which group to split next? 2) How to split this group? 3) How to query the new groups? In the following we will address in detail these challenges.

Which group to split? A good criterion should select a group that has the greatest potential (after the split) to increase class discriminability. Two key factors that influence the split efficiency are: 1) group size and 2) the group impurity, which reflects the mixing of the class labels in the current group. We propose to select the group with the largest expected label information gain if we *were* to split it. More specifically for every leaf group G, we generate a hypothetical split (to be explained next) to form two subgroups G_1 and G_2, and then estimate how much gain the two new groups will lead to. More formally, assuming the label of G is μ, we want to estimate the joint distribution of μ_1 and μ_2, which reflects all possible soft-label combinations one can assign to G_1 and G_2, and then use it to calculate the global **E**xpected label information **G**ain (EGain):

$$EGain(G) = \frac{|G|}{|\mathcal{U}|} Entropy(G, \mu) - \mathbb{E}_{(\mu_1,\mu_2)}\{ \frac{|G_1|}{|\mathcal{U}|} Entropy(G_1, \mu_1) + \frac{|G_2|}{|\mathcal{U}|} Entropy(G_2, \mu_2)\} \quad (2)$$

where $|G|$ denotes the number of instances contained in G, and $Entropy(G, \mu) = -\mu log\mu - (1 - \mu)log(1 - \mu)$. Please note, that by normalizing the gain calculations with respect to $|\mathcal{U}|$ we can compare the gains for groups of different size. Now an open question is: what is the distribution of (μ_1, μ_2)? We approximate this distribution by estimating the distribution of the number of positive and negative instances in each subgroup. Assuming the number of positive instances in each group follows a binomial distribution, then the expected number of positives in the parent group G is $|G|\mu$. Let $n^+ = |G|\mu$ and $n^- = |G|(1 - \mu)$, and let n_1^+ and n_2^+ be the number of positive instances in subgroups G_1 and G_2 respectively. Then $n_1^+ + n_2^+ = n^+$ should hold for all possible combinations of (n_1^+, n_2^+). Thus the distribution of (n_1^+, n_2^+) is a typical hypergeometric distribution, where the probability $P(n_1^+, n_2^+)$ is:

$$P(n_1^+, n_2^+) = \frac{\binom{n^+}{n_1^+}\binom{n^-}{|G_1|-n_1^+}}{\binom{|G|}{|G_1|}} \quad (3)$$

Note that $n_2^+ = n^+ - n_1^+$. Once (n_1^+, n_2^+) is known we can estimate (μ_1, μ_2) using the maximum likelihood approach as $\hat{\mu}_i = n_i^+/|G_i|$ for $i = 1, 2$. Therefore we can approximate the distribution of $P(\mu_1, \mu_2)$ using the hypergeometric distribution in Equation (3), that is, $P(\mu_1, \mu_2) \approx P(n_1^+, n_2^+)$. This approximation can be used to calculate the expected gain in Equation (2). At the end, we choose the group with the largest expected gain from among the candidate groups to perform the split. We note that the calculation of the expected group gain is independent of expected gains for other groups, so the calculation needs to be done only once when the group is labeled.

How to split the chosen group? Unlike decision tree splitting, we do not have any instance-level labels to guide the process. However, there is a simple and often very effective heuristic used frequently in semi-supervised learning: similar data instances tend to have similar class labels (Zhu, Lafferty, and Ghahramani 2003). In other words, dissimilar data tend to carry different labels, and hence, we should try to split the group apart as much as possible.

Following this intuition, our group splitting method aims to find a good attribute d_* along with a splitting value $s^*_{d_*}$ such that the sub-groups generated in this way are most *different* from each other. To achieve this goal, we first find the most diverse dimension d_* that carries a highest data variance:

$$d_* = \arg\max_{d_j \in \mathcal{F}} \mathbb{D}(\{\mathbf{x}_{id_j} | \mathbf{x}_i \in G^*\}) \quad (4)$$

where $\mathbb{D}(.)$ is the variance operation on some set.

The next step is to select a good splitting value $s^*_{d_*}$. In order to split the data apart, we can simply perform 2-means clustering on the values that the data project on attribute d_*. Formally, $s^*_{d_*}$ is given by:

$$s^*_{d_*} = \arg\min_{s \in V} \{ \sum_{v \in V \;\&\&\; v \le s} (v - m_1)^2 + \sum_{v \in V \;\&\&\; v > s} (v - m_2)^2 \} \quad (5)$$

where $V = \{\mathbf{x}_{id_*} | \mathbf{x}_i \in G^*\}$, and m_1, m_2 are the two mean centers corresponding to low and high attribute values respectively.

Updating Soft Labels of the New Sub-Groups After the new sub-groups (sub-populations), denoted by G_1^* and G_2^*, are generated, we need to assess their soft labels. A trick here is that we do not need to query a human to obtain labels for both groups. Instead because there is a probability constraint among the parent group and the child groups, only one child group needs to be queried for label while the other can be calculated by the constraint. Formally, suppose the labels of G^*, G_1^*, G_2^* are μ, μ_1 and μ_2 respectively and say μ_1 is queried from the human by showing the group description which is represented as a conjunctive pattern (e.g. $(2 <= d_j < 12 \wedge d_{j'} = 1)$). Then we can infer:

$$\mu_2 = (\mu \times |G^*| - \mu_1 \times |G_1^*|) \; / \; |G_2^*| \tag{6}$$

3.3 Learning Classifiers from Soft Labeled Groups

So far our group learning algorithm has created a hierarchy of groups, each annotated by a soft label. The question now is how one can build a classifier from the soft-labeled groups of instances. The most straightforward solution is to use instance-based learning algorithms that permit instance weighting or soft labels (Nguyen, Valizadegan, and Hauskrecht 2011).

More specifically, each instance $\mathbf{x}$ in a group G with a soft label $L_G = \mu$ can be represented by two identical instances, one labeled as *Positive* and the other labeled as *Negative*, with respective weights as μ and $1 - \mu$. The model learning then proceeds by training the model from data that include instance duplicates and corresponding labels. Example classifiers that can be trained out of our hierarchy are decision trees, logistic regression, support vector machines and Naive Bayes models.

4 Experiments

In this section we perform an empirical study to evaluate our proposed approach on 12 UCI (Asuncion and Newman 2007) binary classification data sets.

4.1 Methods Tested

We compare our **G**roup-**B**ased **A**ctive **L**earning (GBAL) method to six different baseline methods: two of them are instance-based algorithms: the classic uncertainty sampling (US) (Lewis and Catlett 1994) and random sampling (RS); two are cluster-based algorithms: density-weighted uncertainty sampling (DWUS) (Rashidi and Cook 2011) and hierarchical sampling[1](HS); the fifth one is RIQY which is the state-of-the-art group-based active learning approach with soft-label group queries; and the last one (R-GBAL) implements our group-based active learning framework using random selection of groups instead of heuristics.

4.2 Data Sets

We experiment with 12 UCI binary classification data sets to compare the methods. The datasets come from a variety of real life fields:

- **Physics**: *HIGGS, Seismic-bumps, Ionosphere*
- **Chemistry**: *Biodegradation, Ozone Detection*
- **Medicine**: *Indian Liver Patient Database, Diabetic Retinopathy Debrecen Dataset (Messidor), Blood Transfusion*
- **Website**: *Spambase, Online News Popularity*
- **Life**: *Geographical Origin of Music, Wine Quality*

Some datasets have been used in the study of RIQY (*wine, ionosphere*); some are high-dimensional (*Biodeg, Ozone, Spam, News, Music*); some are unbalanced in class (*Seismic, Ozone, Liver, Transfusion*) and the rest were picked randomly from the UCI repository. The statistics of each data set is listed in Table 1.

4.3 Experimental Settings

Label Assessment While it is easy to simulate the human labeler of instance-based algorithms, for group-based methods, we follow the work of RIQY: given the group description which is a hypercube region, collect all training data that fall into this hypercube and finally report the frequency of occurrence of the two classes in the region. In real life people may not be good at giving very accurate frequency number. For this reason we simulate the soft labels by approximately rounding each frequency number to be in 21-scale on [0,1], i.e. each label can only take one of the values in $\{$0.0, 0.05, 0.1, 0.15,..., 0.95, 1.0$\}$.

Evaluation Metrics To evaluate the methods we split each data set into three disjoint parts: the initially labeled dataset (about 1%-2% of data), a test dataset (about 25% of data) and an unlabeled dataset $\mathcal{U}$ (the rest of data) that is used in the learning phase. Initially labeled data are required by instance-based methods and RIQY. Our methods, GBAL and R-GBAL, do not require any initially labeled set.

The evaluation metric used in the experiment is Area Under the Receiver Operating Characteristic curve (AUC). To reduce the experiment variations all results are averaged over 20 runs in different splits.

Dataset	# of Data	# of features	Major Class	Feature Type
HIGGS	5000	28	53%	Num
Seismic	2584	18	93%	Num-Cat
Ionosphere	351	34	64%	Num
Biodeg	1055	41	67%	Num-Cat
Ozone	1847	72	93%	Num
Liver	579	10	71%	Num-Cat
Messidor	1151	19	53%	Num-Cat
Transfusion	748	5	76%	Num
Spam	4601	57	60%	Num-Cat
News	5000	59	52%	Num-Cat
Music	1059	68	53%	Num
Wine	4898	11	67%	Num

Table 1: 12 UCI data sets. ('Num' and 'Cat' denote Numerical and Categorical feature respectively.)

[1]www.cs.columbia.edu/ djhsu/

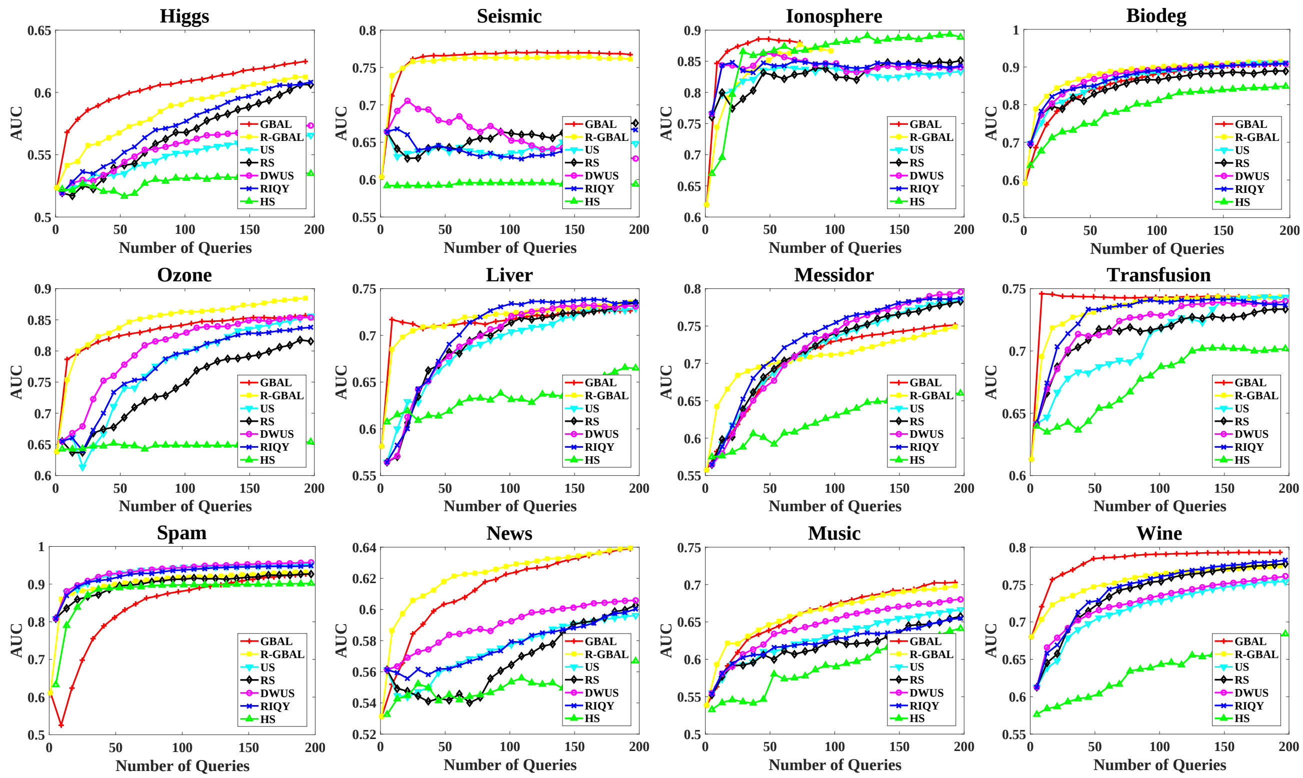

Figure 1: Performances of the different methods on 12 UCI data sets.

4.4 Results

We test all methods by training linear Support vector machines based on the Liblinear package (Fan et al. 2008).

The main results are shown in Figure 1. The graphs plot the AUC quality of the classifiers learned after k queries. We see that our group based method outperforms other baselines on the majority of datasets and is close to the best performing methods on the remaining datasets.

The essential benefit of our approach is that the classifier is able to quickly (after only a very few queries) achieve a good and steady performance. This can be attributed to the fact that our group-based learning strategy is initially trained with high-level supervision, rather than several specific and complicated examples which may easily bias instance-based active learning strategies. The group-based feedback appears very informative especially for high-dimensional data sets or those with unbalanced class distributions.

As the learning process proceeds our method is still leading other methods. It mainly benefits from our active learning strategy which is capable of splitting the *right* groups *appropriately*, in contrast to R-GBAL, which blindly selects groups to perform random splits. In this way, the labels of groups can be quickly refined, with the soft labels of the contained instances being updated multiple times to be more accurate. As a consequence, our model is able to reach a better performance with fewer number of queries. Therefore, it appears, one group query may be more informative than one instance query and our method leverages this information more effectively than other group based methods.

Dataset	RIQY	GBAL
HIGGS	66%	**76%**
Seismic	**71%**	61%
Ionosphere	**83%**	81%
Biodeg	**83%**	72%
Ozone	86%	**89%**
Liver	**60%**	30%
Messidor	**74%**	57%
Transfusion	**36%**	26%
Spam	**81%**	60%
News	73%	**83%**
Music	86%	**88%**
Wine	**58%**	42%

Table 2: The average feature reduction of group queries.

Query Complexity In addition to classification performance, we also compared the complexity of queries generated by the two group based methods: RIQY and GBAL. To make comparison we calculated average feature reduction rate according to Equation (1). The results are shown in Table 2. We can see that RIQY is better at describing groups in a simpler way, however, our model is not very far behind, and at the same time it reaches much better classification performance. Furthermore, this table suggests that it is often not necessary to know every attribute information to answer each group query. This property is crucial to learn from high-dimensional datasets and it can dramatically improve the efficiency of human-computer interaction.

5 Conclusions & Future Work

We have developed a new active learning framework based on group queries. Our framework is good for complex classification tasks that rely heavily on human supervision and their data attributes are easier to describe in groups in contrast to individual examples. The experimental results show two important properties of our current implementation: first by applying group querying and soft labeling technique, we are able to successfully overcome the limitation of the traditional instance-based active learning methods that usually perform poorly when there is just very little supervision; second our active learning splitting heuristic can rapidly refine the groups and their soft labels, thus accelerate the whole model training process.

While our early results are very promising, we would like to note that soft label estimates that are acquired from human annotators may not and typically are not perfect. Hence one important open problem to study is the effect of noisy soft label assessments on the learning process and methods of mitigation of their negative influences.

Acknowledgement

The work presented in this paper was supported by NIH grants R01GM088224 and R01LM010019. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201:81–105.

Asuncion, A., and Newman, D. 2007. Uci machine learning repository.

Bodó, Z.; Minier, Z.; and Csató, L. 2011. Active learning with clustering. *Journal of Machine Learning Research* 16:127–139.

Chakraborty, S.; Balasubramanian, V.; and Panchanathan, S. 2015. Adaptive batch mode active learning. *IEEE transactions on neural networks and learning systems* 26(8):1747–1760.

Dasgupta, S., and Hsu, D. 2008. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, 208–215. ACM.

Druck, G.; Settles, B.; and McCallum, A. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 81–90. Association for Computational Linguistics.

Du, J., and Ling, C. X. 2010. Asking generalized queries to domain experts to improve learning. *Knowledge and Data Engineering, IEEE Transactions on* 22(6):812–825.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9.

Guo, Y., and Schuurmans, D. 2008. Discriminative batch mode active learning. In *Advances in neural information processing systems*, 593–600.

Haque, M. M.; Holder, L. B.; Skinner, M. K.; and Cook, D. J. 2013. Generalized query-based active learning to identify differentially methylated regions in dna. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 10(3):632–644.

Hoi, S. C.; Jin, R.; Zhu, J.; and Lyu, M. R. 2006. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, 417–424. ACM.

Lewis, D. D., and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*.

Nguyen, H. T., and Smeulders, A. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 79. ACM.

Nguyen, Q.; Valizadegan, H.; and Hauskrecht, M. 2011. Learning classification with auxiliary probabilistic information. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, 477–486. IEEE.

Rashidi, P., and Cook, D. J. 2011. Ask me better questions: active learning queries based on rule induction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 904–912. ACM.

Salmani, K., and Sridharan, M. 2014. Multi-instance active learning with online labeling for object recognition. In *The Twenty-Seventh International Flairs Conference*, 406–411.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-instance active learning. In *Advances in neural information processing systems*, 1289–1296.

Settles, B. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6(1):1–114.

Small, K.; Wallace, B.; Trikalinos, T.; and Brodley, C. E. 2011. The constrained weight space svm: learning with ranked features. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 865–872.

Urner, R.; Wulff, S.; and Ben-David, S. 2013. Plal: Cluster-based active learning. In *Conference on Learning Theory*, 376–397.

Valizadegan, H.; Nguyen, Q.; and Hauskrecht, M. 2013. Learning classification models from multiple experts. *Journal of biomedical informatics* 46(6):1125–1135.

XUE, Y., and Hauskrecht, M. 2017. Active learning of classification models with likert-scale feedback. In *SIAM Data Mining Conference, 2071*. SIAM.

Zhu, X.; Lafferty, J.; and Ghahramani, Z. 2003. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3.