Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

New Balanced Active Learning Model and Optimization Algorithm

Xiaoqian Wang[1], **Yijun Huang**[2], **Ji Liu**[2], **Heng Huang**[1*]
[1] Department of Electrical and Computer Engineering, University of Pittsburgh, PA 15261, USA
[2] Department of Computer Science, University of Rochester, NY 14627, USA
xqwang1991@gmail.com, huangyj0@gmail.com, jliu@cs.rochester.edu, heng.huang@pitt.edu

Abstract

It is common in machine learning applications that unlabeled data are abundant while acquiring labels is extremely difficult. In order to reduce the cost of training model while maintaining the model quality, active learning provides a feasible solution. Instead of acquiring labels for random samples, active learning methods carefully select the data to be labeled so as to alleviate the impact from the redundancy or noise in the selected data and improve the trained model performance. In early stage experimental design, previous active learning methods adopted data reconstruction framework, such that the selected data maintained high representative power. However, these models did not consider the data class structure, thus the selected samples could be predominated by the samples from major classes. Such mechanism fails to include samples from the minor classes thus tends to be less "representative". To solve this challenging problem, we propose a novel active learning model for the early stage of experimental design. We use exclusive sparsity norm to enforce the selected samples to be (roughly) evenly distributed among different groups. We provide a new efficient optimization algorithm and theoretically prove the optimal convergence rate $O(1/T^2)$. With a simple substitution, we reduce the computational load of each iteration from $O(n^3)$ to $O(n^2)$, which makes our algorithm more scalable than previous frameworks.

1 Introduction

In many machine learning applications, acquiring labels involves high consumption of time and money. For example, in biomedical application, it requires the intensive work of experts to accurately label a biological specimen. In speech recognition, acquiring annotation of words in speech is a human-intensive process which calls for a long period of work by trained linguists [Zhu *et al.*, 2005]. In order to reduce the cost of label collection while maintaining the learning model quality, active learning methods were proposed to select a small portion of data for labeling from a large pool of candidates such that the model constructed with the labeled data has the optimal potential performance.

In general, there are three different settings of scenarios in active learning: 1) membership query synthesis; 2) stream-based selective sampling; and 3) pool-based sampling. Membership query synthesis automatically generates samples for labeling, where the quality of the arbitrarily generated instance is not guaranteed. For example, in a handwriting recognition task [Baum and Lang, 1992], many query images generated by the model do not even contain an identifiable character, which makes it hard to label manually. Stream-based selective sampling treats data in a sequential manner, where the model checks the data one by one from the data source and decides whether to label the current sample or not. However, such model could not leverage the connection between candidate data, thus the query decision tends to be biased and sensible when the input distribution is not known. As for pool-based sampling, it treats unlabeled data as a whole and draws a small bunch of data to label, such that certain metric can be maximized. Detailed survey of active learning can be found in [Settles, 2010; Guyon *et al.*, 2011; Monteleoni, 2006].

In active learning, many previous methods select only one instance to label in each iteration. However, such mechanism fails to take the correlation among multiple instances into consideration thus loses information. On the contrary, selecting in the batch mode [Guo and Schuurmans, 2008] is more favorable since the information overlap between data samples can be utilized in the learning process.

According to the selective standard, active learning methods can be roughly divided into two categories: one type of methods focus on picking out the most "informative" data, *i.e.,* select the data with the maximum entropy [Dagan and Engelson, 1995] or least confidence [Culotta and McCallum, 2005]. On the other hand, some methods tend to find the most "representative" data with the highest representative power of the data structure [Yu *et al.*, 2006; Nie *et al.*, 2013], such as clustering structure, manifold, *etc.* The methods in latter category is more desired for the early stage of experimental design when the number of labeled data is limited. In this paper we focus on this type of models for the early stage of active learning.

In order to find the representative data in active learning,

*Corresponding author.

previous methods tried to find a small set of data such that all unlabeled data could be represented as a linear combination of the selected ones [Yu *et al.*, 2006; Nie *et al.*, 2013]. However, such model did not employ the class structure among samples, which could make the selected samples to be predominated by the major classes. Such mechanism does not make good use of information from the minor classes thus loses corresponding representative power. To deal with this problem, in this paper we propose a new active learning model which uses exclusive sparsity norm as regularization and enforce a structural sparsity such that the selected samples are (roughly) evenly distributed among different groups. This setting guarantees that the selected samples contain information from all groups of data, thus are more "representative". We provide a new efficient optimization algorithm and theoretically prove the optimal convergence rate $O(1/T^2)$. To the best of our knowledge, this is the first optimization algorithm with such convergence rate for general minimization problems with exclusive sparsity norm.

We conducted experiments on 8 benchmark datasets and evaluated the performance of our method. We verified that our method constructed a fairly good model by labeling just a few samples. Such results validated the potential of our model to relieve the heavy burden of labeling samples.

2 Balanced Active Learning Selection

In early stage of experimental design, given a set of n unlabeled data $X = [\mathbf{x}_1,\ \mathbf{x}_2, ...,\ \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, the goal of active learning is to select m ($m < n$) data points from the set to label, such that the model constructed with the m labeled data has the maximized potential performance. Such task can be formulated as the following optimization problem as is done in transduction experimental design (TED) [Yu *et al.*, 2006]:

$$\min_{V\in\mathbb{R}^{m\times n}, B\subset X, |B|=m} \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - B\mathbf{v}_i\|^2 + \lambda \|\mathbf{v}_i\|^2, \quad (1)$$

where $V = [\mathbf{v}_1,\ ...,\ \mathbf{v}_n]$, the constraints $B \subset X$ and $|B| = m$ indicate that matrix B contains a subset of m data from the total set of n data in X. However, Problem (1) is NP-hard. In [Yu *et al.*, 2006], the original problem is approximated with sequential optimization problem, which leads to an inefficient optimization method. To deal with this, in [Nie *et al.*, 2013] and [Cong *et al.*, 2011], the researchers proposed the following problem:

$$\min_{W\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - X\mathbf{w}_i\|^2 + \lambda \|W\|_{2,1}, \quad (2)$$

where $\|W\|_{2,1} := \sum_{i=1}^{n} \|W_i\|$ (W_i is the ith row of W).

Naturally, the value of $\|W_i\|$ indicates the importance of $\mathbf{x}_i$ (the i-th sample in X), thus the subset of data can be selected based on W. The motivation of Model (2) is to approximate $\|W\|_{2,0}$ by the $\ell_{2,1}$-norm, such that the structured sparsity is enforced on the W matrix. This model selects a subset of samples which could denote all other samples with linear combination, hence guarantees the representative and informative power of the selected samples.

However, all the above methods did not consider the potential class structure. It is possible that all selected samples are taken up by the predominant group, such that the information from the minor groups are ignored. To tackle this challenging problem, we propose a new balanced active learning model with exclusive sparsity norm, which enforces the selected samples for labeling to be (roughly) evenly distributed among different groups.

For a vector $\mathbf{v} \in \mathbb{R}^n$, its exclusive sparsity norm (ℓ_e-norm) is defined as: $\|\mathbf{v}\|_e := \sqrt{\sum_{g\in\mathcal{G}} \|\mathbf{v}_g\|_1^2}$, where $\mathcal{G}$ is a hyper set consisting of k disjoint index sets (each index set g is a subset of $\{1, 2, \cdots, n\}$), and $\mathbf{v}_g$ is the sub-vector of $\mathbf{v}$ indexed by g. The exclusive sparsity norm regularization serves to enforce a structural sparsity on the solution such that its nonzeros are (roughly) evenly distributed in different groups.

Our new active learning objective function is to solve:

$$\min_{W\in\mathbb{R}^{n\times n}} \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - X\mathbf{w}_i\|_2^2 + \lambda \|W\|_{2,e}^2, \quad (3)$$

where $\|W\|_{2,e} := \sqrt{\sum_{g\in\mathcal{G}} \|W_g\|_{2,1}^2}$ and W_g denotes a sub matrix of W with rows in W indexed by g. In active learning, we don't have label information. Thus, we can conduct clustering (such as K-means) on the data to get k groups, where k is the number of classes.

In our new model, the $||W||_{2,e}$ regularization term forces the structured sparsity $\ell_{2,1}$-norm within each group, hence the representative samples in each group are selected. Meanwhile, the exclusive sparsity norm imposes ℓ_2-norm between groups, *i.e.* the squared norms of each group are summed together. As a result, all groups are considered as evenly important and the representative samples are selected from all groups without suppressing individual group. Thus, our new active learning model selects the representative samples for labeling with incorporating the intrinsic group structure information.

3 Optimization

Our previous iterative reweighted algorithms [Nie *et al.*, 2010; Gao *et al.*, 2015] can be applied to solve the new objective, but there is no convergence rate guarantee. Thus, we will derive a new efficient optimization algorithm, with the guarantee of the optimal convergence rate $O(1/T^2)$.

3.1 Preliminary Rules

For optimization problems with the structure: "smooth loss $F(\mathbf{x})$ + non-smooth function $H(\mathbf{x})$"

$$\min_{\mathbf{x}} \quad F(\mathbf{x}) + H(\mathbf{x}), \quad (4)$$

the Nesterov type of accelerated algorithms [Nesterov, 2007], for example, FISTA [Beck and Teboulle, 2009] in Algorithm 1, is often considered as one of the most efficient gradient based algorithms, since it achieves the optimal convergence rate $O(1/T^2)$ where T is the number of iterations. $H(\mathbf{x})$ could be the regularizer such as $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|$ or the indicator function representing the constraint, for example, $\mathbf{I}_{\mathbf{x}\geq 0}(\mathbf{x})$. One can see the key step in FISTA is Step 3 (the

proximal step) in Algorithm 1, which essentially solves the following problem:

$$\text{Prox}_{\gamma H(\mathbf{x})}(\mathbf{c}) := \arg\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{c}\|^2 + \gamma H(\mathbf{x}) \qquad (5)$$

where $\mathbf{c} = \bar{\mathbf{x}} - \gamma\nabla F(\bar{\mathbf{x}})$. The steplength γ can be set as any positive constant smaller than $(\max_x \|\nabla^2 F(x)\|)^{-1}$ or decided by using the linear search scheme [Beck and Teboulle, 2009]. The difficulty of solving (4) is decided by the complexity of $H(\mathbf{x})$. When $H(\mathbf{x})$ is simple enough such as $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|$, or $\mathbf{I}_{\mathbf{x}\geq 0}(\mathbf{x})$, the proximal step (5) is easy. However, if $H(\mathbf{x})$ is $\|\mathbf{x}\|_e$, it is difficult to solve efficiently. Therefore we need reformulation work to simplify the proximal steps. The following preliminary results play the key roles in reformulation and solving proximal steps.

Algorithm 1 The General Framework of FISTA

Require: $\mathbf{x}^{\text{old}}$, γ (step length)
Ensure: $\mathbf{x}^{\text{new}}$
1: Initialize $t^{\text{old}} = 1$, $\bar{\mathbf{x}} = \mathbf{x}^{\text{old}}$;
2: **while** not converge **do**
3: $\quad \mathbf{x}^{\text{new}} = \text{Prox}_{\gamma H(\mathbf{x})}(\bar{\mathbf{x}} - \gamma\nabla F(\bar{\mathbf{x}}))$;
4: $\quad t^{\text{new}} = \frac{1}{2} + \frac{1}{2}\sqrt{1+4t^{\text{old}}}$;
5: $\quad \bar{\mathbf{x}} = \mathbf{x}^{\text{new}} + (\mathbf{x}^{\text{new}} - \mathbf{x}^{\text{old}})(t^{\text{old}} - 1)/t^{\text{new}}$;
6: $\quad \mathbf{x}^{\text{old}} = \mathbf{x}^{\text{new}}$, $t^{\text{old}} = t^{\text{new}}$;
7: **end while**

Lemma 1. *Algorithm 2 exactly solves $\mathbf{P}_1^{\zeta}(\mathbf{a}, b)$ – the projection of $[\mathbf{a}; b]$ onto the ℓ_1 norm cone $\{[\mathbf{x}; y] \mid \|\mathbf{x}\|_1 \leq y\}$:*

$$\mathbf{P}_1^{\zeta}(\mathbf{a}, b) := \arg\min_{\mathbf{x}\in\mathbb{R}^d, y} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \frac{\zeta}{2}(y-b)^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq y. \qquad (6)$$

Problem (6) can be considered as a general version of the projection onto the ℓ_1 cone. Although (6) does not have the closed form, but its solution can be obtained from a search routine in Algorithm 2. Algorithm 2 essentially gives a method to search a feasible point satisfying the KKT condition. The complexity of this algorithm is $O(d \log d)$. The key motivation behind this algorithm is the monotonicity of the optimal solution $\mathbf{x}^*$, that is, if $|\mathbf{a}_i| \geq |\mathbf{a}_j|$, then $|\mathbf{x}_i^*| \geq |\mathbf{x}_j^*|$.

Lemma 1 follows the strategy of projection on to the ℓ_1 norm cone, which can be considered as a special case with $\zeta = 1$. The proof mainly applies fundamental results such as KKT condition in optimization. We include their proofs for completeness.

Proof. Let us consider a trivial case first: $\|\mathbf{a}\|_1 \leq b$. In this case, the optimal values for $\mathbf{x}$ and y are $\mathbf{a}$ and b respectively.

Then let us consider the nontrivial case: $\|\mathbf{a}\|_1 > b$. In this case, it is easy to see that the optimal values $\mathbf{x}^*$ and y^* satisfy $\|\mathbf{x}^*\|_1 = y^*$ and the element signs of $\mathbf{x}^*$ are the same as $\mathbf{a}$. Therefore, to simplify the following notation and discussion, we make an assumption without the loss of generality

$$\mathbf{a}_1 \geq \mathbf{a}_2 \geq \cdots \geq \mathbf{a}_d \geq 0.$$

Algorithm 2 $[\mathbf{x}, y] = \text{P}_1^{\zeta}(\mathbf{a}, b)$

Require: $\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$, $\zeta \in \mathbb{R}$
Ensure: $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
1: Sort the sequence $\{|\mathbf{a}_j| \mid j = 1, \cdots, d\}$ in the decreasing order and denote $|\mathbf{a}|^{(j)}$ the jth largest element of $\mathbf{a}$ in the absolute value sense and define $|\mathbf{a}|^{(d+1)}$ as 0;
2: $t = 0$;
3: **for** $j = 1 : d$ **do**
4: $\quad t = t + |\mathbf{a}|^{(j)}$;
5: $\quad \delta = (t-b)/(\zeta^{-1} + j)$;
6: $\quad$ **if** $|\mathbf{a}|^{(j+1)} \leq \delta \leq |\mathbf{a}|^{(j)}$ **then**
7: $\quad\quad \mathbf{x} = \text{sgn}(\mathbf{a}) \odot \max(\mathbf{0}, |\mathbf{a}| - \delta)$;
8: $\quad\quad y = \|\mathbf{x}\|_1$;
9: $\quad\quad$ Return;
10: $\quad$ **end if**
11: **end for**
12: $\mathbf{x} = \mathbf{a}$;
13: $y = b$;

Based on this assumption, the problem (6) is equivalent to solving the following problem

$$\min_{\mathbf{x},y} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \frac{\zeta}{2}(y-b)^2 \quad \text{s.t.} \quad \mathbf{1}^\top\mathbf{x} = y,\ \mathbf{x} \geq \mathbf{0}. \qquad (7)$$

It can be further simplified by

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{a}\|^2 + \frac{\zeta}{2}(\mathbf{1}^\top\mathbf{x}-b)^2 \quad \text{s.t.} \quad \mathbf{x} \geq \mathbf{0}. \qquad (8)$$

The optimal solution is defined by the KKT condition

$$0 \leq \mathbf{x}_i \perp \mathbf{x}_i - \mathbf{a}_i + \zeta(\mathbf{1}^\top\mathbf{x} - b) \geq 0 \quad \forall i \in \{1, 2, \cdots, d\}.$$

We also note that if $\mathbf{a}_i \geq \mathbf{a}_j$, then $\mathbf{x}_i^* \geq \mathbf{x}_j^*$ (otherwise we can simply swap the values of $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ to obtain a lower objective value). Therefore, from the monotonicity assumption on $\mathbf{a}$, we have:

$$\mathbf{x}_1^* \geq \mathbf{x}_2^* \geq \cdots \geq \mathbf{x}_d^*.$$

Let $j \in \{1, 2, \cdots, d\}$ be the watershed: $\mathbf{x}_i^* = 0$ for $i > j$ and $\mathbf{x}_i^* \geq 0$ for $i \leq j$. The KKT is simplified by:

$$\forall i \leq j \quad \mathbf{x}_i^* = \mathbf{a}_i - \zeta(\mathbf{1}^\top\mathbf{x}^* - b) \geq 0,$$
$$\forall i > j \quad \mathbf{x}_i^* = 0 \quad \mathbf{a}_i - \zeta(\mathbf{1}^\top\mathbf{x}^* - b) \leq 0.$$

Summarizing all $\mathbf{x}_i^*$'s to obtain:

$$\mathbf{1}^\top\mathbf{x}^* = \left(\sum_{i=1}^{j}\mathbf{a}_i\right) - j\zeta(\mathbf{1}^\top\mathbf{x}^* - b)$$
$$\Rightarrow \mathbf{1}^\top\mathbf{x}^* = \frac{\left(\sum_{i=1}^{j}\mathbf{a}_i\right) + j\zeta b}{1 + j\zeta}$$
$$\Rightarrow \zeta(\mathbf{1}^\top\mathbf{x}^* - b) = \frac{\left(\sum_{i=1}^{j}\mathbf{a}_i\right) - b}{\zeta^{-1} + j} =: \delta_j.$$

Then finding a point satisfying the KKT condition is equivalent to finding a j such that

$$\begin{aligned} &\forall i \leq j \quad \mathbf{x}_i^* = \mathbf{a}_i - \delta_j > 0, \\ &\forall i > j \quad \mathbf{x}_i^* = 0 \quad \mathbf{a}_i - \delta_j \leq 0. \end{aligned} \tag{9}$$

Due to the monotonicity, to find such "j", we only need to enumerate all possible values for j such that

$$\mathbf{a}_j - \delta_j \geq 0, \, \mathbf{a}_{j+1} - \delta_j \quad \leq 0. \tag{10}$$

As long as we find such "j^*", we can compute the optimal values for $\mathbf{x}^*$ from (9)

$$\mathbf{x}^* = \max(\mathbf{0}, \mathbf{a} - \delta_{j^*})$$

and y^* from $y^* = \mathbf{1}^\top \mathbf{x}^*$. If we remove the assumption, the definition of δ_j should be modified into

$$\delta_j := \frac{\left(\sum_{i=1}^{j} |\mathbf{a}|^{(i)}\right) - b}{\zeta^{-1} + j},$$

where $|\mathbf{a}|^{(i)}$ denotes the ith largest absolute value in $\mathbf{a}$. The condition to define the optimal j in (10) should be replaced by

$$|\mathbf{a}|^{(j)} - \delta_j \geq 0, \, |\mathbf{a}|^{(j+1)} - \delta_j \quad \leq 0.$$

The optimal values for $\mathbf{x}^*$ and y^* are given by respectively

$$\mathbf{x}^* = \mathrm{sgn}(\mathbf{a}) \odot \max(\mathbf{0}, |\mathbf{a}| - \delta_{j^*}) \quad y^* = \|\mathbf{x}^*\|_1.$$

Algorithm 2 exactly follows the procedure to find the optimal solution. It completes the proof. □

3.2 Smooth Loss Function + ℓ_e Regularization: FISTA-LOCP Algorithm

We consider the following general formulation

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_e^2. \tag{11}$$

One might ask why do not use $\phi\|\mathbf{w}\|_e$ as the regularizer. The reason lies on that $\frac{\lambda}{2}\|\mathbf{w}\|_e^2$ is easier to solve efficiently and leads to the same solution as using $\phi\|\mathbf{w}\|_e$ if $\lambda(\phi)$ is appropriately chosen. Note that $\|\mathbf{w}\|_e^2$ is still non-smooth.

To directly apply the FISTA framework, one has to solve the proximal step (5) with $H(\cdot) = \frac{\lambda}{2}\|\mathbf{w}\|_e^2$. However, it is very difficult to solve it efficiently in general. Existing approaches apply iterative algorithms to *approximately* solve this proximal step, for example, in [Kong *et al.*, 2014; Yuan and Yan, 2011], thus requiring heavy computation load (computing the inverse of a $n \times n$ matrix) and unable to theoretically ensure the convergence (rate).

We use a simple substitution as below to reformulate the problem (11), which largely simplifies the original formulation.

$$\mathrm{Prox}_{\gamma H(\cdot)}(\mathbf{c}) := \arg\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w} - \mathbf{c}\|^2 + \frac{\gamma\lambda}{2}\|\mathbf{w}\|_e^2. \tag{12}$$

The following will show that this problem can be solved efficiently due to Lemma 1. (If we use $\|\mathbf{w}\|_e$ as the regularizer, then (12) is difficult to solve efficiently.) This why we are interested in (11).

Because groups are disjoint, the proximal step can be split into a few subproblems

$$\min_{\mathbf{w}_g} \quad \frac{1}{2}\|\mathbf{w}_g - \mathbf{c}_g\|^2 + \frac{\gamma\lambda}{2}\|\mathbf{w}_g\|_1^2 \quad \forall g \in \mathcal{G}.$$

To solve this problem, we can reformulate it into the form of (6)

$$\min_{\mathbf{w}_g, y} \quad \frac{1}{2}\|\mathbf{w}_g - \mathbf{c}_g\|^2 + \frac{\gamma\lambda}{2}y^2 \quad \text{s.t.} \quad \|\mathbf{w}_g\|_1 \leq y$$

whose solution is exactly provided by $\mathbf{P}_1^{\gamma\lambda}(\mathbf{c}_g, 0)$.

Now we can simply apply the proximal operator defined in (12) to the FISTA framework to obtain an algorithm with the optimal convergence rate. Since the proximal step in this algorithm is the projection onto the ℓ_1 cone, we call this algorithm as L-one cone projection (FISTA-LOCP) algorithm. One can verify that the computation load per iteration is still on the gradient, that is, $O(n^2)$, if the number of samples is proportional to n.

3.3 Optimization Algorithm for Problem (3)

Here we introduce the steps for applying FISTA-LOCP to Problem (3). According to the discussion above, we use $\|W\|_{2,e}^2$ as the regularization instead and reformulate Problem (3) as:

$$\begin{aligned} &\mathrm{Prox}_{\gamma H(\cdot)}(C) \\ &:= \arg\min_{W} \frac{1}{2}\|W - C\|_F^2 + \frac{\gamma\lambda}{2}\|W\|_{2,e}^2. \end{aligned} \tag{13}$$

where $C = \bar{W} - \gamma\nabla F(\bar{W})$ with $F(W) = \frac{1}{2}\|X - XW\|_F^2$. Problem (13) can be split into a few proximal subproblems below:

$$\min_{W_g} \frac{1}{2}\|W_g - C_g\|_F^2 + \frac{\gamma\lambda}{2}\|W_g\|_{2,1}^2 \quad \forall g \in \mathcal{G}. \tag{14}$$

The subproblem in (14) can be further rewritten as

$$\min \sum_{i=1}^{|g|} \frac{1}{2}\|W_{g_i} - C_{g_i}\|^2 + \frac{\gamma\lambda}{2}\left(\sum_{i=1}^{|g|}\|W_{g_i}\|\right)^2, \tag{15}$$

where g_i denotes the ith element in g, W_{g_i} denotes the g_i-th row of W and C_{g_i} is the g_i-th row of C.

Since

$$\|W_{g_i} - C_{g_i}\|^2 = \|W_{g_i}\|^2 + \|C_{g_i}\|^2 - 2W_{g_i}^T C_{g_i},$$

we can easily prove that Problem (15) can be minimized only when W_{g_i} has the same direction with C_{g_i}. That being the case, what matters is to compute the length of W_{g_i}, thus we deal with the following problem:

$$\min \sum_{i=1}^{|g|} \frac{1}{2}(\|W_{g_i}\| - \|C_{g_i}\|)^2 + \frac{\gamma\lambda}{2}\left(\sum_{i=1}^{|g|}\|W_{g_i}\|\right)^2, \tag{16}$$

Define vector $\mathbf{s} \in \mathbb{R}^{|g|}$ and $\mathbf{t} \in \mathbb{R}^{|g|}$ such that $s_i = \|W_{g_i}\|$ and $t_i = \|C_{g_i}\|$, then Problem (16) can be written as (6)

$$\min_{\mathbf{s}, \mathbf{y}} \quad \frac{1}{2}\|\mathbf{s} - \mathbf{t}\|^2 + \frac{\gamma\lambda}{2}y^2 \quad \text{s.t.} \quad \|\mathbf{s}\|_1 \leq y$$

whose solution is exactly provided by $\mathbf{P}_1^{\gamma\lambda}(\mathbf{t}, 0)$. $\mathbf{s}$ shows the cardinality of W_{g_i}. After we get $\mathbf{s}$ and $\mathbf{t}$, we can recover W_g such that $W_{g_i} = \frac{s_i}{t_i} C_{g_i}$. We summarize the steps for optimizing Problem (3) in Algorithm 3. Still, one can verify that the computation load per iteration is $O(n^2)$, if the number of samples is proportional to n.

In practice, the exclusive sparsity norm introduces row sparsity to the W matrix, which renders the number of non-zero rows in W much smaller than n and thereby decreases the real computational time. Moreover, as we consider different groups independently in Algorithm 3, we can implement the update in a parallel way to adapt to the large scale setting.

Algorithm 3 Algorithm for Problem (3)

Require: $X \in \mathbb{R}^{d\times n}$, $W^{\text{old}} \in \mathbb{R}^{n\times n}$, $\mathcal{G}$, γ (step length)
Ensure: W^{new}
1: Initialize $t^{\text{old}} = 1$, $\bar{W} = W^{\text{old}}$;
2: **while** not converge **do**
3: $C = \bar{W} - \gamma \nabla F(\bar{W})$;
4: **for** $g \in \mathcal{G}$ **do**
5: $t_i = \|C_{g_i}\|$, $i = 1, \cdots, |g|$;
6: $[\mathbf{s}, y] = \mathbf{P}_1^{\zeta}(\mathbf{t}, 0)$;
7: $W_{g_i}^{new} = \frac{s_i}{t_i} C_{g_i}$, $i = 1, \cdots, |g|$;
8: **end for**
9: $t^{\text{new}} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4t^{\text{old}}}$;
10: $\bar{W} = W^{\text{new}} + (W^{\text{new}} - W^{\text{old}})(t^{\text{old}} - 1)/t^{\text{new}}$;
11: $W^{\text{old}} = W^{\text{new}}$, $t^{\text{old}} = t^{\text{new}}$;
12: **end while**

4 Experimental Results

In this section, we conduct experiments to evaluate our method. We select a subset of samples from the training data to quire labels and then construct classification models according to these labeled data. The goal is to pick out the most representative samples such that the constructed classification model maintains high discriminative power.

4.1 Experimental Settings

We compare our method with two baseline methods: 1. **Random** sample selection which arbitrarily selects a certain number of samples from a given set and quire labels; 2. **K-means** sample selection which uses K-means clustering to partition the data into several clusters and select points closest to the K centroids as the candidate query points. Moreover, we compare against two state-of-the-art methods, including: 3. **QUIRE** (QUerying Informative and Representative Examples) [Huang *et al.*, 2010], which picks out the informative and representative instances based on the min-max view of active learning; 4. **RRSS** (Robust Representation and Structured Sparsity) [Nie *et al.*, 2013]. RRSS method is closely related to ours, which imposes $\ell_{2,1}$-norm as a structured sparsity regularization to find representative samples.

In the active learning experiments, we randomly pick out half of the data for training while the other half for testing. We apply different active learning methods to the training data and select a subset to query labels. The selection of data to query is based on the magnitude of the l_2-norm of each row. Then we construct an SVM model based on the queried data. The test data is used to evaluate performance of the constructed classification model.

We adopt the libsvm toolbox [Chang and Lin, 2011] to implement SVM classification and set the hyperparameter of regularization term as 1. We use two-fold cross validation and record the average classification accuracy among the two repetitions. For methods involving hyper-parameters, *i.e.,* λ for QUIRE in Eq. (5) of [Huang *et al.*, 2010], γ for RRSS in Eq. (6) of [Nie *et al.*, 2013], and λ in Eq. (3) of our method, we tune the hyper-parameters in the range of $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$. For K-means clustering, we set the number of clusters as the ground truth. We use 100 random initialization for K-means and retain the best result among these 100 repetitions with respect to K-means objective function value. For our method, we use the K-means clustering result as the group allocation of training data.

We first conduct experiments on 6 relatively balanced benchmark data sets, which include: Aggregation [Gionis *et al.*, 2007], Binalpha[1], Compound [Zahn, 1971], R15 [Veenman *et al.*, 2002], BreastCancer and Seeds. The last two datasets are downloaded from UCI repository [Lichman, 2013]. The evaluation on these 6 benchmark datasets is based on the classification accuracy on the test data. Moreover, we include 2 imbalanced datasets from KEEL-dataset [Alcalá-Fdez *et al.*, 2011], glass0123vs456 and newthyroid1, to evaluate the methods when the data exhibits severe class imbalance. The evaluation on the 2 imbalanced datasets is based on the F1 score on the test data.

4.2 Performance Analysis with Different Number of Query Samples

In the experiment, we assess the performance of the above active learning methods with various number of queried samples. We set the number of queried samples to be $\{k, 2k, \cdots, 5k\}$, respectively, where k is the number of classes in each data set. We can notice from the data description in the previous subsection that k is a very small value comparing to the number of samples. Our goal is to construct a good classification model with only a limited number of samples labeled, which is of great applicable importance when acquiring labels causes much effort.

The performance comparison is summarized in Table **??** and **??**. It shows that our method performs equally or even better than all other methods on all data sets. By comparing our method with the K-means method, we notice an apparent increase of classification accuracy which indicates that directly using points closest to the centroid may not be the best choice. Instead, our method manages to pick out the more "representative" data points in each cluster by applying the grouped $\ell_{1,2}$-norm regularization. If we compare the RRSS result with our method, we can find the superiority of considering data cluster structure in active learning. RRSS imposes $\ell_{2,1}$-norm as the sparse regularization in the model, such that a few "important" data can be found in the learning process. However, RRSS doesn't take the group information among

[1] http://www.cs.nyu.edu/ roweis/data.html

	Data Sets	Random	K-means	QUIRE	RRSS	FISTA-LOCP
k Queried Samples	Aggregation (k=7)	0.5444	0.7817	0.8477	0.4391	**0.8541**
	BreastCancer (k=2)	0.4786	0.6501	0.7555	0.3499	**0.8829**
	Binalpha (k=36)	0.2023	0.1517	0.2094	0.1610	**0.2707**
	Compound (k=6)	0.5240	0.6415	**0.6843**	0.3088	0.6543
	R15 (k=15)	0.4317	0.3883	0.4750	0.4000	**0.9783**
	Seeds (k=3)	0.4810	0.4524	0.7571	0.7905	**0.9048**
2k Queried Samples	Aggregation (k=7)	0.8033	0.7944	0.8553	0.4391	**0.8832**
	BreastCancer (k=2)	0.7436	0.6501	0.8974	0.3499	**0.9223**
	Binalpha (k=36)	0.3191	0.1610	0.3063	0.2429	**0.3860**
	Compound (k=6)	0.5965	0.7268	0.7095	0.4963	**0.7943**
	R15 (k=15)	0.5033	0.5950	0.5750	0.4667	**0.7300**
	Seeds (k=3)	0.6048	0.6667	0.8476	0.8571	**0.8810**
3k Queried Samples	Aggregation (k=7)	0.8198	0.8477	0.8604	0.5520	**0.8794**
	BreastCancer (k=2)	0.9019	0.6501	**0.9546**	0.3499	0.9531
	Binalpha (k=36)	0.4031	0.1652	0.3932	0.3063	**0.4694**
	Compound (k=6)	0.6966	0.7268	0.6516	0.6518	**0.8245**
	R15 (k=15)	0.6100	0.6167	0.6400	0.4667	**0.7167**
	Seeds (k=3)	0.7381	0.6762	0.8714	0.8857	**0.9048**
4k Queried Samples	Aggregation (k=7)	0.8731	0.8579	0.8896	0.5508	**0.9086**
	BreastCancer (k=2)	0.9312	0.8785	0.9576	0.3499	**0.9619**
	Binalpha (k=36)	0.4608	0.1667	0.4274	0.3654	**0.5235**
	Compound (k=6)	0.7345	0.7242	0.7295	0.6969	**0.8270**
	R15 (k=15)	0.6500	0.6933	0.6717	0.4667	**0.8000**
	Seeds (k=3)	0.8429	0.7190	0.8905	**0.9000**	0.8857
5k Queried Samples	Aggregation (k=7)	0.8871	0.8731	0.8947	0.5508	**0.9124**
	BreastCancer (k=2)	0.9312	0.9092	**0.9722**	0.3499	0.9619
	Binalpha (k=36)	0.5078	0.1667	0.4872	0.4167	**0.5491**
	Compound (k=6)	0.6820	0.7367	0.7169	0.6969	**0.7541**
	R15 (k=15)	0.7467	0.6917	0.6933	0.4667	**0.8817**
	Seeds (k=3)	0.8476	0.7381	0.9000	0.9000	**0.9095**

Table 1: Classification accuracy comparison with different number of samples queried. In the table, k means the number of clusters in each data set. The best result in each setting is marked in boldface.

	Data Sets	Random	K-means	QUIRE	RRSS	FISTA-LOCP
3k Queried Samples	glass0123vs456 (k=2)	0.6290	0.6350	0.5295	0.6222	**0.6667**
	newthyroid1 (k=2)	0.6686	0.5127	0.7471	0.3685	**0.7576**
4k Queried Samples	glass0123vs456 (k=2)	0.5892	0.6148	0.7659	0.6339	**0.8286**
	newthyroid1 (k=2)	0.7396	0.3543	0.6548	0.3148	**0.7585**
5k Queried Samples	glass0123vs456 (k=2)	0.7817	0.6824	0.7607	0.7386	**0.8596**
	newthyroid1 (k=2)	0.7396	0.4574	0.8725	0.7016	**0.8948**

Table 2: F1 score comparison on two imbalanced datasets. In the table, k means the number of clusters in each data set.

data points into consideration, which may cause imbalance in the selection of data (*i.e.,* the majority of selected data come from a few predominant clusters, while data points in other clusters are left out.) In contrast, our method tend to select a few data from each cluster, which enhances the "representation" power of the selected subset.

Moreover, we can observe an interesting phenomenon that the superiority of our method tends to be more obvious when the number of queried samples is smaller. It indicates that our model is capable of finding a representative subset with querying only a few samples, which saves the heavy burden for label collection.

5 Conclusion

In this paper, we proposed a novel active learning model with exclusive sparsity norm. Unlike previous ones, our model considers the group structure among samples such that the information from both major and minor groups can be incorporated in active learning. Such mechanism makes the selected samples to be more "representative". We propose an efficient optimization algorithm and theoretically prove the optimal convergence rate $O(1/T^2)$. We evaluate our model on 8 benchmark datasets and find good classification performance with few queried samples.

Acknowledgements

This work was partially supported by the following grants: NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, NSF-CCF1718513, NIH R01 AG049371.

References

[Alcalá-Fdez *et al.*, 2011] Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

[Baum and Lang, 1992] Eric B Baum and Kenneth Lang. Query learning can work poorly when a human oracle is used. In *International Joint Conference on Neural Networks*, volume 8, page 8, 1992.

[Beck and Teboulle, 2009] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Cong *et al.*, 2011] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. *CVPR*, pages 3449–3456, 2011.

[Culotta and McCallum, 2005] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

[Dagan and Engelson, 1995] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 150–157. The Morgan Kaufmann series in machine learning,(San Francisco, CA, USA), 1995.

[Gao *et al.*, 2015] Hongchang Gao, Weidong Cai, and Heng Huang. Anatomical annotations for drosophila gene expression patterns via multi-dimensional visual descriptors integration. *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2015)*, 2015.

[Gionis *et al.*, 2007] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):4, 2007.

[Guo and Schuurmans, 2008] Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pages 593–600, 2008.

[Guyon *et al.*, 2011] Isabelle Guyon, Gavin C Cawley, Gideon Dror, and Vincent Lemaire. Results of the active learning challenge. *Active Learning and Experimental Design@ AISTATS*, 16:19–45, 2011.

[Huang *et al.*, 2010] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, pages 892–900, 2010.

[Kong *et al.*, 2014] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding. Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. *NIPS*, 2014.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[Monteleoni, 2006] Claire E Monteleoni. Learning with online constraints: shifting concepts and active learning. 2006.

[Nesterov, 2007] Y. Nesterov. Gradient methods for minimizing composite objective function. *Technical Report*, 2007.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint l2,1-norms minimization. *NIPS*, 2010.

[Nie *et al.*, 2013] Feiping Nie, Hua Wang, Heng Huang, and Chris HQ Ding. Early active learning via robust representation and structured sparsity. In *IJCAI*, 2013.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[Veenman *et al.*, 2002] Cor J Veenman, Marcel JT Reinders, and Eric Backer. A maximum variance cluster algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(9):1273–1280, 2002.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM, 2006.

[Yuan and Yan, 2011] X. Yuan and S. Yan. A finite newton algorithm for non-degenerate piecewise linear systems. *AISTAT*, pages 841–854, 2011.

[Zahn, 1971] Charles T Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *Computers, IEEE Transactions on*, 100(1):68–86, 1971.

[Zhu *et al.*, 2005] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, language technologies institute, school of computer science, 2005.