A Structured Perspective of Volumes on Active Learning

Xiaofeng Cao

Advanced Analytics Institute, University of Technology Sydney
xiaofeng.cao@student.uts.edu.au

Abstract

We approximate the version space which covers all feasible classification hypotheses into a structured geometric hypersphere against agnostic distribution. We present a structured perspective that divides the available active learning (AL) sampling approaches into two kinds of strategies: *Outer Volume Sampling* and *Inner Volume Sampling*. For the outer volume, it is represented by a circumscribed hypersphere which would exclude any outlier (non-promising) hypothesis from the version space globally. While for the inner volume, it is represented by many inscribed hyperspheres, which cover most of hypotheses within the outer volume. To enhance the performance of AL, we aggregate the two kinds of volumes to eliminate their sampling biases via finding the optimal inscribed hyperspheres in the enclosing space of outer volume. We then propose a Volume-based Model for the AL sampling without any distribution assumption. To generalize our theoretical model, in a non-linear feature space, spanned by kernel, we use sequential optimization to globally optimize the original space to a sparse space by halving the size of the kernel space. Then, the EM (Expectation Maximization) model which returns the local center helps us to find a local representation. To describe this process, we propose an *easy-to-implement* algorithm called Volume-based AL (VAL). Empirical evaluation on a various set of structured clustering and unstructured handwritten digit data sets have demonstrated that, employing our proposed model can accelerate the decline of the prediction error rate with fewer sampling number compared with the other algorithms.

Keywords: Active learning, version space, hypothesis, hypersphere, outer volume, inner volume.

1. Introduction

In many real-world applications, collecting adequate training inputs with the annotation help of the domain experts is often expensive and time consuming. This motivates the key idea of Active Learning (AL) [1], which interactively queries the labels of unlabeled instances to minimize the training outputs with human supervisions. In such a sampling scenario, the AL algorithms update the current classification hypothesis by accepting the label

annotation on a single or group of unlabeled data. It is thus most of the AL approaches are a kind of hypothesis-based strategy. After adding new instances to the training set, the learning algorithm prefers to query those instances which maximize the "*distance*" between the current and updated classification hypotheses. Therefore, AL is a supervised learning task and benefits the work in text processing [2], image annotation [3], multi-label classification [4] [5], etc. Since the learner samples the instances strategically, the number of training outputs to learn a strong learning hypothesis can often be much smaller than the number required in a standard supervised learning. However, the labeled data often are available but inadequate in real applications, and how to minimize the amount of prior labeled data to reduce the dependence of input hypothesis (prior labels) [6] still remains to be studied.

In theoretical study, active learners usually use a notion termed *version space* [7] to model the generalization performance of any practical learning algorithm. In their assumptions, the hypothesis class which includes all feasible hypotheses is mapped into a bounded convex body, and shrinking the *volume* of this convex body by a given cut size can help the learner to find the optimal hypothesis. For example, in version space, the hypothesis which can maximize the hypothesis distance between the current and updated hypothesis is the primary sampling object. Correspondingly, in real-world, a single or group of unlabeled data which can largely update the current training model should be picked up. Moreover, a common policy whatever in theoretical or practical level is greedy search, i.e., the learner need to consider all candidate hypotheses or unlabeled data to produce the next hypothesis update. In real-world AL tasks, the approach which heuristically searched the whole unlabeled pool to evaluate which data was the most highly informative, had attracted the attention of learners, and this approach was called "*uncertainty sampling*". However, the cost of this greedy strategy is expensive. To study more effective model, in theoretical level, [6] utilized the approach of convex optimization to approximate the version space to a hyperellipsoid. In such a geometrical body, it tightly [8] encloses the most of the hypotheses. Then, they cut the hyperellipsoid into a half-space that included any instance whose class label could not evidently be inferred from the hypothesis trained so far, rather than focusing upon maximal uncertainty instances. Although it has attracted the attention of learners, the hyperellipsoid still has been primarily of theoretical interest since there is not enough evidences are discovered to convince us in the infinite dimension space. Different from it, the hypersphere has obtained more provable guarantees in version space description, such as [9] [10] [11] [12], etc.

In this paper, these evidences motivate us to use the hypersphere to approximate the version space in high dimension space. By transferring the AL sampling issue into a hypothesis update process in version space theory, we observe there are two criteria for the AL sampling process: maximizing the hypothesis update, and minimizing an enclosing set with high representation to the version space. The former takes the "highly informative" [6] data as the sampling targets, and while the latter considers the representative data as the sampling targets.

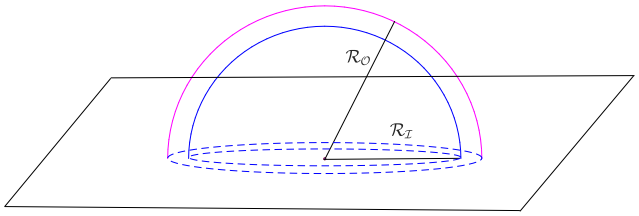
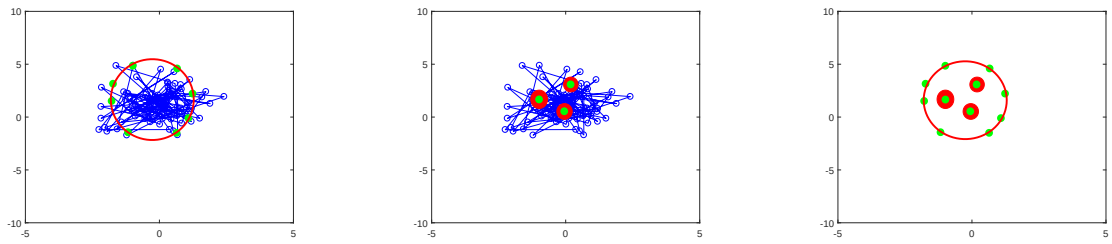


Figure 1: This illustration shows the fresh perspective of approximating the version space as an enclosing sphere, and we show one half-space of it in three dimensional space, in which the enclosing space of the hemisphere with a radius of $\mathcal{R}_{\mathcal{I}}$ represents the *Inner Volume* of the half-space, and the enclosing space between it and another hemisphere with a radius of $\mathcal{R}_{\mathcal{O}}$ represents the *Outer Volume* of the half-space.

Interestingly, the hypotheses that nearest to the optimal classification hypothesis lie on the surface of the version space, and the hypotheses that have high similarity to its local space lie inside the version space. To describe these earlier AL studies, we take a fresh perspective on them and define the surface part of the version space as *Outer Volume* and the internal part of the version space as *Inner Volume*, respectively (see Figure 1). Meanwhile, we define the informativeness evaluation approaches of AL as *Outer Volume Sampling* strategy, and the representation sampling approaches of AL as *Inner Volume Sampling* strategy. Without a specified distribution assumption, shrinking the volume of version space become a smart solution to model the sampling progress over an agnostic distribution.

However, the optimal performance of one classification learning model is not easy to obtain, such as no learners know which hyperplane of SVM classifier is the best, although some of them obtain high accuracies on the prediction results. Usually, machine learning community tries to train a ϵ-optimal hypothesis with finite VC dimension, where $\epsilon \leq 1$ [13] [14] [15]. Therefore, the optimal hypothesis is outside the version space and may have multiple possible positions which surrounds the version space (see Theorem 1 and Remark 2). To circumvent the limitation of this uncertainty, the outer volume is represented by the surface of the version space (see Figure 2(a)), which excludes any outlier (non-promising) hypotheses from the version space globally, and the inner volume is represented by many inscribed hyperspheres, which cover all feasible hypotheses within the outer volume (see Figure 2(b)). Since the AL based on the two kinds of volumes may have sampling biases in terms of noises, overlapping classes, and local convergence, we use both of them to represent the version space to ignore the non-promising hypothesis globally and cover all local hypothesis locally (see Figure 2(c)). To obtain this structured representation, we find the optimal representation inscribed hyperspheres in the enclosing space



(a) Outer Volume (b) Inner Volume (c) Structured representation

Figure 2: This group of illustrations show the motivation of this paper, where each blue point denotes one classification hypothesis, and the green points lie in the outer or inner volumes represent a target hypothesis for AL. (a) The outer volume is denoted by a circumscribed hypersphere which will exclude any outlier (non-promising) hypotheses from the version space globally. (b) The inner volume is represented by many inscribed hyperspheres, which covers all feasible hypothesis within the outer volume. (c) Structured representation by outer and inner volumes provides a succinct representation to ignore the non-promising hypothesis globally and covers all local hypothesis locally.

of outer volume, in which each hypersphere is represented by its local hypersphere center. As described in this representation sampling process, we propose a theoretical framework called *Volume-based AL Model*.

To generalize this theoretical framework in the real world AL tasks, we firstly use the transductive experimental design of statistics regression to globally map the data space to a sparse space which excludes all outlier hypotheses and shrinks the number of candidate sampling set into a half. After obtaining the sparse structure of data space, the Expectation Maximization (EM) model which returns local centers can provide an effective local representation to the enclosing set of outer volume, i.e., the current inner volume of data space. Finally, we propose the Volume-based Active Learning (VAL) algorithm. Contributions of this paper are described as follows:

- We approximate the version space into a hypersphere and divide it into two parts: outer volume and inner volume.
- We provide a theoretical guarantee for dividing the earlier AL approaches into two kinds: *Outer Volume Sampling* and *Inner Volume Sampling*.
- We design a theoretical AL framework termed "Volume-based AL Model" in the version space, which consists of the outer and inner volumes to find an optimal representation for version space globally and locally.
- To generalize this theoretical description, we provide an *easy-to-implement* algorithm called VAL (Volume-based Active Learning) in Euclidean space.

- The proposed VAL algorithm is dependent of classifier and prior labels which results in a faster error rate decline, compared with other AL approaches.

The remainder of the paper is structured as follows. Section 2 presents the related work including AL in version space, and descriptions of outer volume and inner volume. Section 3 then introduces the notions of version space and we divide the available AL models into two categories of strategies. In Section 4, we present the motivation of this paper by discussing the relationship between the AL models and volumes in version space. In Section 5, we propose a Volume-based AL Model which is a theoretical framework in version space. To implement it, we propose a Volume-based AL algorithm by sequential and expectation maximization optimizations in Section 6. The experiments are reported in Section 7. The discussions are presented in Section 8. We conclude this paper in Section 9.

2. Related Work

Active learners tend to select the informative instances that split the version space into two parts, in which the external part contains the sparse examples that lie on the surface of the version space, called *Outer volume* of the version space. The internal part contains most volume of the version space, called *Inner volume*.

To present our fresh perspective, Section 2.1 describes the AL in structured version space which contains all feasible classification hypotheses, then Section 2.2 and 2.3 explain the outer and inner volume sampling in AL, respectively.

2.1. Active learning in version space

Learning a hypothesis from labeled instances is not a universally applicable paradigm [16]. Many natural learning tasks involved with sampling new unlabeled examples are not simply passive, but instead make use of at least some form of AL strategies to examine the proposed problem domain. By active learning, any form of learning task can have some control over the inputs on which it trains. Then, the sampling outputs using the greedy learning strategy become possible.

In such learning problems, Mitchell [7] described the learning task based on the partial ordering of original inputs in version space. It required the learners to do active learning by examining the sampled target instances whether fall in the "difference" hypothesis regions. Before learning a new hypothesis, learners firstly examine the information already given and then evaluate the uncertainty of a candidate region. To reduce the *label complexity*, a series approaches of partitioning version space were proposed, in which a theoretical foundation was that the objective learning function could be perfectly expressed by one hypothesis in the version space. Under this policy, reducing the volume of version space becomes a theoretical description for AL.

However, calculating the volume of a convex body is hard because of the computationally intractable [17]. To study the target hypothesis distribution, we observe that the highly informative hypothesis are far away the most hypotheses, and they are distributed on the surface of the convex body, and a highly representative hypothesis is distributed in a dense internal region of the convex body.

In the geometric approximation description of machine learning, Enclosing Cylinder (EC), Enclosing Ball (EB), Minimum-Width Annulus (MWA) [18] are the three spatial geometry description tools. Of them, EB attracts most attention and obtains provable guarantees. By extracting the core sets [19] [20] that "represents" the data space, [21] [22] [23] [24] [25] have utilized EB to improve the performances of SVM and clustering in high dimensional space. In addition to this, the surprising properties of hypersphere are independent of the dimension and have been widely used in gap tolerant classifiers [26], KNN search, 1-cylinder problem [18], sphere trees [27] and so on. Therefore, we use the hypersphere to describe the version space.

2.2. Outer volume of version space

There are two fundamental propositions in AL theory: (1) maximizing the hypothesis update by iterative sampling, and (2) representation sampling. Usually, the hypothesis or local distribution that farthest to the current hypothesis or distribution lie on the surface of the version space. Therefore, the sampling targets of these labeled-based AL approaches lie on the *Outer Volume* of version space and these labeled-based AL sampling are called *Outer Volume Sampling*.

For the outer volume sampling, lack of rich prior experience transfered learners' attention on the available labeled resource and then motivated the learning approaches of pool-based AL [28]. In this learning framework, they selected the unlabeled data independently from the candidate pool via observing their underlying hypothesis or distribution update after querying. As one of the important pool-based model, [9] designed a relevance feedback strategy that measured the uncertain class assignments of unlabeled data in each iteration. The idea of iterative sampling then was used in [29], [5], [12], [30], etc., which set the unlabeled data into a pool to wait for picking out based on a given sampling strategy trained by the current classification hypothesis or labeled data. However, the error rate curve could not decline significantly in case of a very small amount of sampling number or labeled data.

2.3. Inner volume of version space

Different from outer volume, representation sampling is to optimize an effective mapping structure for the original version space and the potential learning rule is to keep the *diameters* of arbitrary local spaces whatever in hypothesis or distribution metric. Therefore, the sampled hypothesis must lie inside the version space and this type of approach is called *Inner Volume Sampling*.

Compared with the outer volume sampling, clustering-based approaches is one branch of inner volume sampling since it studies the clustering structure to get help from the hypothesis that lie inside the version space. For example, [31] actively labeled the credible sub clustering trees with root node's label by a probability discriminant model. While it showed negative AL performances in the terms of clustering result with high error rate, unstructured data space, and so on. An important potential reason was lack of labeled data for querying and then led to an inevitable poor performance in AL tasks. Such situation also appeared in references of [32], [33], [34], etc. But actually these approaches were still labeled-based.

Moreover, less support from the labeled data may lead to the performance decrease of AL, and the need of the priori label amount at the beginning of training is seriously underestimated. Besides it, there are many existing AL approaches that could not be adopted well in a learning task with insufficient amount of labels, such as Margin [9], Hierarchical[31], Quire[35], Re-active[36], [12]. Considering to reduce the dependence to label amount, [37], [38], [39], [40], [41], [42] used the approach of representation sampling to map the original version space. To keep a low loss mapping, [12] measured the diameter of the version space and then mapped a representation space with the similar space diameter. However, they neglected the importance of local metric in space mapping process.

3. Version Space and Active Learning Strategies

We are *the first* to propose the fresh perspective of considering the AL as finding the most informative or representative hypothesis from the huge version space which covers all possible hypothesis. This theoretical description aims to improve the reliability for any possible AL framework by volume.

As we observe that the current AL sampling targets can lie on the surface or in the internal of version space, we divide the version space into two parts: "Outer Volume" and "Inner Volume". In this section, Section 3.1 describes the version space, and Section 3.2 divides the AL into two kinds of strategies, where the used main notations are described in Table 1.

3.1. Version space

Consider a data space $\mathcal{X}$ with n points $\{x_1, x_2, ..., x_n\}$, a distribution assumption $\mathcal{D}$ over $\mathcal{X}$, and a classification hypothesis set $\mathcal{H}$ with finite VC dimension, where $x_i = [x_{i1}, x_{i2}, ..., x_{im}]$, $\mathcal{H} = \{h_1, h_2, ..., h_k\}$.

Assumption 1. *Some theoretical descriptions of version space are based on the parameter space of classification hyperplane in Euclidean space. Our definition is based on the VC dimension. Since there are no specified description about this notation, we will learn from some characteristics of Euclidean space in this paper.*

Table 1: A summary of notations

Notation	Definition
$\mathcal{X}$	data space/set
x_i	the ith data point of $\mathcal{X}$
$\mathcal{D}$	distribution assumption over $\mathcal{X}$
$\mathcal{H}$	hypothesis set over $\mathcal{X}$
h_i	the ith hypothesis of $\mathcal{H}$
$n, k, \mathcal{K}, \epsilon, \kappa$	constants
E	hypothesis space
$\ell(\cdot, \cdot)$	metric function
$Vol(\cdot)$	the geometric volume of the input object
d	diameter of E
h^*	the optimal hypothesis
$h_{<\cdot>}$	a hypothesis with special setting
$\mathcal{O}$	outer volume of $\mathcal{H}$
$\mathcal{I}$	inner volume of $\mathcal{H}$
$B(\cdot, \cdot)$	the enclosing ball with special radius and center settings
$\mathcal{R}_{<\cdot>}$	radius of hypersphere with special setting
$\mathcal{C}_{<\cdot>}$	center of hypersphere with special setting
B', B^*	enclosing balls with special settings
$\mathcal{S}^+, \mathcal{S}^-$	half-spaces of $\mathcal{H}$
$Pr(\cdot\|\cdot)$	conditional probability
$\theta_{<\cdot>}$	vector angle with special setting
$\mathcal{L}$	loss function
$\mathcal{A}, \mathcal{B}$	matrices

Definition 1. *Version space [7] [6].* The graph G which connects all possible hypothesis is the version space,

and it is an ordered triple,

$$G = \{V, E, \ell\} \tag{1}$$

and

$$\begin{aligned} V_i &= h_i, \forall\, i = 1, 2, ..., n \\ E_{ij} &= \{h_i, h_j\}\ i, j \in (1, k) \end{aligned} \tag{2}$$

where E represents the hypothesis space, E_{ij} is the edge of ith and jth hypothesis, and ℓ is the metric function. In this graph, any two vertices have an edge and their distance metric is defined as follows.

Definition 2. *Hypothesis distance* [12] [43] [44]. Given hypothesis h_i and h_j $(i, j \in (1, k))$, the distance between them is:

$$\ell(h_i, h_j) = \{h_i(x_i) \neq h_j(x_i), \forall\, i = 1, 2, ..., k\} \tag{3}$$

where $\ell(\cdot, \cdot)$ denotes the distance between the two inputs.

Definition 3. *Diameter of the version space* [43] [44]. The edge with the maximum hypothesis distance of E denotes the diameter of the version space, that is to say,

$$d = \underset{i,j \in (1,k)}{argmax}\{E_{ij}\} \tag{4}$$

where d denotes the diameter of E.

3.2. Active learning strategies

In the data space $\mathcal{X}$ with n samples, the hypothesis number of querying k data is $\mathcal{C}_n^\rho$. However, no learner knows how to obtain the optimal hypothesis h^*. Here we discuss the hypothesis number of classifying m classes:

Theorem 1. *The VC dimension [13] [14] [15] of $\mathcal{H}$ is about 2^n, i.e., there are approximately 2^n hypotheses in the version space.*

Proof. Assume the querying number $\rho \leq \mathcal{K} \leq n$ in a ρ-class setting, here we obtain:

$$\begin{aligned} VC(\mathcal{H}) &= \mathcal{C}_n^\rho + \mathcal{C}_n^{\rho+2} + \cdots + \mathcal{C}_n^n \\ &= 2^n - \sum_{i=1}^{\rho-1} \mathcal{C}_n^i \\ &= 2^n - \varepsilon \end{aligned} \tag{5}$$

where ε is a constant. ■

Remark 1. *This theorem shows the solution of an arbitrary classification issue is "*unclosed-form*" although the hypothesis could lead to a perfect classification accuracy. In the goal of the advanced AL theory, the learner would try to produce an "*closed-form*" sampling set which is independent on classifier category and parameter space.*

Therefore, we have (the following remark will be used in *Lemma 7*)

Remark 2. *The optimal hypothesis is not contained in the version space, that is to say*

$$h^* \notin \mathcal{H} \tag{6}$$

Definition 4. Active Learning. AL sampling helps to minimize the difference between the optimal hypothesis and the final hypothesis, that is to say

$$\min_{h_f} \ \ell(h_f, h^*) \tag{7}$$

where h_f represents the classification hypothesis trained on the final labeled set after sampling, and $h^* \notin \mathcal{H}$.

Generally, the learners iteratively sample the data point which can maximize the hypothesis update in the version space. Then, we have

Definition 5. *Active Learning Sampling.* Let $\mathcal{C}_0$ be the initialization training set, AL is to find the data $\Phi(x)$ which changes the current hypothesis greatly:

$$\underset{\Phi(x)}{argmax} \ \ell(h_{\mathcal{C}_0}, h_{\mathcal{C}_0'}) \tag{8}$$

where $h_{\mathcal{C}_0}$ represents the current classification hypothesis, $h_{\mathcal{C}_0'}$ represents the updated hypothesis after adding $\Phi(x)$ to training set $\mathcal{C}_0$, and $\mathcal{C}_0' = [\mathcal{C}_0 \ \Phi(x)]$.

From the above definition, we highlight two AL strategies corresponding to hypothesis update and representation sampling:

Strategy 1. *Maximizing the hypothesis update*. Learners should identify pairs of hypothesis in the hypothesis space E with maximum distance,

$$E' = \{\{h_1, \hat{h}_1\}, \{h_2, \hat{h}_2\}, ..., \{h_k, \hat{h}_k\}\} \tag{9}$$

where $\ell(h_i, \hat{h}_i) \leq \ell(h_i, h_j), \forall i, j \in (1, k)$, and it is used in [9], [29], [45], [5], [12], [30], etc.

Strategy 2. *Representation sampling*. Minimizing a sub version space of $\mathcal{D}'$ which is similar with $\mathcal{D}$, that is to say,

$$\ell(\mathcal{D}, \mathcal{D}') \to 0 \tag{10}$$

where $\mathcal{D}' \subset \mathcal{D}$, and this strategy is used in [46], [35], [47], [48], etc.

4. Motivation

AL theory studies the classification hypothesis (Section 2.1) issue via iteratively sampling a data which can maximize the hypothesis update (Strategy 1) or minimizing a sub set with high representation to the original space (Strategy 2). Observing the two strategies, we find the Strategy 1 favours to to sample the hypothesis lie on the surface of version space since they are close to the optimal hypothesis, but Strategy 2 tends to select the hypothesis lying inside the version space since the local representation is the default sampling rule. Therefore, Strategy 1 is the AL approach based on *Outer Volume Sampling*, and Strategy 2 is the AL approach based in *Inner Volume Sampling*.

To prove this perspective, this section will discuss their potential distributions of the sampling targets of the two different strategies to support our division, where Section 4.1 describes the volume of version space and divides the volume of the version space into two parts−outer and inner volume, and Section 4.2 presents our perspective involved with the distribution of the target hypothesis of AL sampling. Then, Section 4.3 and 4.4 present theoretical understanding on this perspective.

4.1. Volumes of version space

Volume is a theoretical notion for the size of the version space. To describe this high dimensional space, we approximate it to a hypersphere, and divide it into two parts: outer volume and inner volume. In this section, we show that the relationship between the two kinds of volumes.

Let $\mathcal{O}$ and $\mathcal{I}$ represent the outer and inner volume respectively, here we remark:

Remark 3. *The geometric volume of the version space is the volume sum of $\mathcal{O}$ and $\mathcal{I}$, i.e.,*

$$Vol(\mathcal{O}) + Vol(\mathcal{I}) = Vol(\mathcal{H}) \tag{11}$$

where $Vol(\cdot)$ [19] denotes the geometric volume of the input objective.

Assume $\mathcal{O}$ and $\mathcal{I}$ can be described as the MEB (Minimum EB) issues of $B(\mathcal{R}_{\mathcal{O}}, \mathcal{C}_{\mathcal{O}})$ and $B(\mathcal{R}_{\mathcal{I}}, \mathcal{C}_{\mathcal{I}})$, we remark

Remark 4. *$\mathcal{O}$ and $\mathcal{I}$ are two concentric hyperspheres which satisfy*

$$\begin{aligned} \mathcal{R}_{\mathcal{O}} &= (1+\epsilon)\mathcal{R}_{\mathcal{I}} \\ \mathcal{C}_{\mathcal{O}} &= \mathcal{C}_{\mathcal{I}} \end{aligned} \tag{12}$$

where ϵ is an infinitesimal constant.

We then need the following theorems to understand the outer and inner volume.

Theorem 2. *Let $\mathcal{Z}$ be the largest hypersphere contained $\mathcal{X}$, then,*

$$\mathcal{R}_{\mathcal{I}} \leq \mathcal{R}_{\mathcal{Z}}/(m(1+\epsilon)) \tag{13}$$

Proof. To obtain the upper bound on $\tau = \mathcal{R}_{\mathcal{I}}/\mathcal{R}_{\mathcal{V}}$, we consider the volumes of $\mathcal{Z}$ and $\mathcal{O}$. The plainly

$$\frac{\mathcal{R}_{\mathcal{O}}^m}{\mathcal{R}_{\mathcal{Z}}^m} \leq \frac{Vol(\mathcal{O})}{Vol(\mathcal{Z})} \tag{14}$$

Following [19] [49], they have proved τ is $\frac{1}{m}$, thus

$$\frac{\mathcal{R}_{\mathcal{I}}^m}{\mathcal{R}_{\mathcal{Z}}^m} \leq \frac{1}{(m(1+\epsilon))^m} \tag{15}$$

and so

$$\frac{\mathcal{R}_{\mathcal{I}}}{\mathcal{R}_{\mathcal{Z}}} \leq \frac{1}{d(1+\epsilon)} \tag{16}$$

as stated. ■

Theorem 3. *Assume that the ball $\mathcal{O}$ is exactly tight. For any closed half-space B' that contains the center $\mathcal{C}_{\mathcal{O}}$ in $B(\mathcal{R}_{\mathcal{O}}, \mathcal{C}_{\mathcal{O}})$, it must contain at least one point from $\mathcal{H}$ that is at distance $\mathcal{R}_{\mathcal{O}}$ from the center $\mathcal{C}_{\mathcal{O}}$.*

Proof. Let us use B represent $B(\mathcal{R}_{\mathcal{O}}, \mathcal{C}_{\mathcal{O}})$, and suppose there exist κ points in the closed half-space B'. Then, there must have a point $h_{\kappa'}$ which satisfies

$$\mathcal{R}_{\mathcal{I}} \leq \ell(h_{\kappa'}, \mathcal{C}_{\mathcal{O}}) = \mathcal{R}_{\mathcal{O}} \tag{17}$$

Otherwise, $Vol(B') < Vol(B)/2$, and there will exist a tighter MEB B^* for containing $\mathcal{H}$, i.e., $\exists B^*$ which satisfies $Vol(B^*) < Vol(B)$. The theorem follows. ■

4.2. Active learning by volumes

Considering the hypothesis with the maximum distance to the current hypothesis is located in the outer volume of the version space, and the representation hypothesis is located inside the version space, i.e., the inner volume, we present our structured perspective as below.

Theorem 4. *The hypothesis that farthest to the current hypothesis must lie in the outer volume of the version space, that is to say,*

$$\hat{h}_i \in \mathcal{O} \; \forall i = 1, 2, ..., n \tag{18}$$

where $\mathcal{O}$ represents the hypothesis set of the outer volume.

Theorem 5. *h_c lies in the inner volume of the version space, that is to say,*

$$h_c \in \mathcal{I} \tag{19}$$

where h_c represents the center hypothesis of an arbitrary inscribed hypersphere, and $\mathcal{I}$ represents the hypothesis set of inner volume of the version space.

4.3. Outer volume sampling

To prove *Theorem 4*, we need the following lemmas to discuss the upper and lower bounds of $\ell(\hat{h}_i, h_i)$ and $\ell(h^*, h_t)$, where h_t represents the target hypothesis when observing the underlying distribution of h^*.

Lemma 6. *The bound of $\ell(\hat{h}_i, h_i)$ is $\sqrt{||\ell(h_i, h_t)||^2 - \mathcal{R}_{\mathcal{O}}^2} \leq \ell(\hat{h}_i, h_i) \leq 2\mathcal{R}_{\mathcal{O}}$ s.t. $\overrightarrow{h_t h_{\mathcal{O}}} \cdot \overrightarrow{h_i h_{\mathcal{O}}} = 0$, where $h_t \in \mathcal{O}$, and $h_{\mathcal{O}}$ represents the hypothesis of $\mathcal{C}_{\mathcal{O}}$.*

Proof. Upper bound. Suppose that the diameter through h_i is d', here we divide d' into two parts: d'^+ and d'^-, where we set $h_i \in d'^-$. Based on the characteristics, we have the following results:

$$\begin{aligned} 0 \leq \ell(h_i, h_t) \leq \mathcal{R}_{\mathcal{O}}, \forall h_t \in \mathcal{S}^- \\ \mathcal{R}_{\mathcal{O}} \leq \ell(h_i, h_t) \leq 2\mathcal{R}_{\mathcal{O}}, \forall h_t \in \mathcal{S}^+ \end{aligned} \tag{20}$$

where $\mathcal{S}^-$ and $\mathcal{S}^+$ represents the half-space which contains d'^- and d'^+, respectively. Then the upper bound of $\ell(\hat{h}_i, h_i)$ as stated.

Lower bound. Suppose the vector $\overrightarrow{h_\Gamma h_\Lambda}$ tangent to $\overrightarrow{d'}$, where $h_\Gamma, h_\Lambda \in \mathcal{O}$, then we have $\mathcal{R}_{\mathcal{O}}^2 + ||\ell(h_{\mathcal{O}}, h_i)||^2 = ||\ell(h_t, h_i)||^2 +, \forall h_t \in \{h_\Omega, h_\Lambda\}$. Then, the lower bound of the lemma follows. ■

Lemma 7. *As claimed of Theorem 4,* $\ell(\hat{h}_i, h_i)_{max} = \ell(h_{\mathcal{O}}, h_i) + \mathcal{R}_{\mathcal{O}}$.

Proof. Following Lemma 6, $\hat{h}_i \in \mathcal{S}^+$. Give an arbitrary hypothesis $h_\Omega \in \mathcal{S}^+$, then by the triangle inequality, we have

$$\ell(h_\Omega, h_i) \leq \ell(h_{\mathcal{O}}, h_i) + \mathcal{R}_{\mathcal{O}} \tag{21}$$

Here we find when $sin < \overrightarrow{h_{\mathcal{O}} h_\Omega}, \overrightarrow{h_i h_{\mathcal{O}}} >= 0$ we have the maximum of $\ell(\hat{h}_i, h_i)$, where $\hat{h}_i = h_\Omega \in \mathcal{O}$. Then the lemma follows. ■

Lemma 8. *Let* h_t *be the target hypothesis and* h^**, the bound of* $\ell(h_t, h^*)$ *is* $\ell(h^*, h_\Psi) \leq \ell(h_t, h^*) \leq 2\mathcal{R}_{\mathcal{O}} + \ell(h^*, h_\Psi)$*, where* h_Ψ *is the nearest intersection between* $\overrightarrow{h^* h_t}$ *and* ***E****.*

Proof. By the triangle inequality, we have

$$\ell(h_t, h_\Psi) - \ell(h^*, h_\Psi) \leq \ell(h_t, h^*) \leq \ell(h^*, h_\Psi) + \ell(h_t, h_\Psi) \tag{22}$$

When $\ell(h_t, h_\Psi) = \mathcal{R}_{\mathcal{O}}$, $\ell(h^*, h_\Psi) + \ell(h_t, h_\Psi)$ will have $\ell(h_t, h^*)_{max} = 2\mathcal{R}_{\mathcal{O}} + \ell(h^*, h_\Psi)$. When $h_t = h_\Psi$, we will have $\ell(h_t, h^*)_{min} = \ell(h^*, h_\Psi)$. Therefore, lemma 8 follows. ■

4.4. Inner volume sampling

To prove Theorem 5, we need to discuss why the target hypothesis h_t are distributed inside the local hypersphere. In the following propositions, we use the probability distribution by taking different hypotheses as priori observation hypothesis, and then we set the MMD metric as distribution measurement to further explain our perspective.

Proposition 1. *Assume the local hypersphere B' has infinite hypotheses with uniform distribution,* h_Γ *is distributed on the surface of B' and* h_Λ *is located inside B'. Let* h_t *be an arbitrary hypothesis in B', we can find* $\sum_{h_t \in B'} Pr_{h_t \in B'}(h_t | h_\Gamma) \leq \sum_{h_t \in B'} Pr_{h_t \in B'}(h_t | h_\Lambda)$.

Proof. Let us discuss the bounds of $\ell(h_t, h_\Gamma)$ and $\ell(h_t, h_\Lambda)$: $0 \leq \ell(h_t, h_\Gamma) \leq 2\mathcal{R}'$, and $0 \leq \ell(h_t, h_\Gamma) \leq 2\mathcal{R}'$, where $\mathcal{R}'$ is the local radius of B'. Assume the distance metric matrices of h_Γ and h_Λ to B' respectively are $\mathcal{A}$ and $\mathcal{B}$, we can find $\mu_(\mathcal{A}) \leq \mu(\mathcal{B})$ and $\sigma_(\mathcal{A}) \leq \sigma(\mathcal{B})$. Then, the lemma is as stated and h_c should be located inside B'. ■

Proposition 2. *Let MMD be the distribution metric, then the distribution distance of representation hypothesis and original local hypothesis ball meets:* $\ell(h_\Gamma, B') > \ell(h_\Lambda, B')$.

Proof. Let $\mathcal{F}$ be a class of functions $f\colon \mathcal{X} \to \mathbb{R}$, and X and Y are two distributions with m' and n' samples, respectively. Then, the distance between the two distributions is estimated as:

$$MMD(\mathcal{F}, X, Y) := \underset{f \in \mathcal{F}}{sup}(\frac{1}{\mathbf{m}} \sum_{i=1}^{m'} f(x_i) - \frac{1}{\mathbf{n}} \sum_{i=1}^{n'} f(y_i)) \tag{23}$$

Since $\ell(h_\Gamma, B') = \ell(h_\Gamma, h_{B'}) > \ell(h_{B'}, h_\Gamma)$, we have $MMD(h_\Gamma, B') > MMD(h_\Lambda, B')$. Then, the lemma is as stated and h_c should be close to the center of B'. ■

As the outer volume of the version space covers most of the hypotheses and we have set ϵ as infinitesimal constant, the internal part of the local hypersphere should be located inside the outer volume. Then, Theorem 5 follows.

5. Theoretical Active Learning Model

Active learner of outer volume has formal the guarantees that hold when the approximated MEB (Minimum Enclosing Ball) of the version space is separable with margins. To implement this assumption, one would need to exclude all the outlier hypotheses. Returning to greedy selection of the outer volume in the version space, we could see that the underlying distribution over hypotheses that could not provide a margin-dependent approximation guarantee without labeled hypothesis as prior experience. Therefore, finding the optimal inscribed hyperspheres could reduce the dependence to labeled hypothesis.

In this section, a volume-splitting strategy termed Volume-based AL Model is presented to find the optimization representation for the original version space, where Section 5.1 claims the motivation of this volume-splitting strategy, Section 5.2 presents the methodology of excluding the outlier hypotheses, Section 5.3 describes the finding process of the optimal inscribed hyperspheres, and Section 5.4 proposes the Volume-based AL Model.

5.1. Motivation of volume-splitting

As we claimed, the version space is a theoretical approximation of data filed of Euclidean space. Therefore, we would discuss the relationship between the underlying distribution of the classification hyperplane by training the data of outer volume and inner volume in Euclidean space. Here we present our perspective

Theorem 9. *Hypothesis set lies in the outer volume of the class is the subset of its inner volume, that is to say,*

$$\mathcal{H}_\mathcal{O} \subset \mathcal{H}_\mathcal{I} \tag{24}$$

where $\mathcal{H}_\cdot$ represents the hypothesis set of the input object.

To prove this theorem, we discuss it in settings of non-crossed MEBs and crossed MEBs in the following lemmas.

Lemma 10. *Let $\theta_{\mathcal{O}}$ and $\theta_{\mathcal{I}}$ be the angles between the classification hyperplane and outer volume, inner volume, respectively. For a pair of non-crossed MEBs, the angle range of $\theta_{\mathcal{O}}$ is smaller than that of $\theta_{\mathcal{I}}$.*

Proof. Given β_1^+ and β_2^- are one pair of data points which has the smallest distances in Non-crossed MEBs, i.e., it satisfies the following assumption:

$$\begin{aligned} &\ell(\beta_1^+, \beta_2^-) < \ell(\beta_i^+, \beta_j^-), for\ all\ i, j \\ s.t.\ &B^+ = \{\beta_1^+, \beta_2^+, ..., \beta_\eta^+\}, B^- = \{\beta_1^-, \beta_2^-, ..., \beta_{\eta'}^-\} \end{aligned} \tag{25}$$

where η and η' are the data number of the MEB B^+ and B^-, respectively. Suppose $\overrightarrow{W}$ be the parameter vector of the classification hyperplane h_w, ν be the intersection point of h_w and $\overrightarrow{\beta_1^+, \beta_2^-}$, $\overrightarrow{\mathcal{C}_{\mathcal{O}} P_1}$ and $\overrightarrow{\mathcal{C}_{\mathcal{O}} P_2}$ be two vectors in the MEB with maximum volume, and

$$\begin{aligned} &\overrightarrow{W} \perp \overrightarrow{\mathcal{C}_{\mathcal{O}} P_1}\ and\ \overrightarrow{W} \perp \overrightarrow{\mathcal{C}_{\mathcal{O}} P_2} \\ s.t.\ &||\mathcal{C}_{\mathcal{O}} - P_1||_2 = \mathcal{R}_{\mathcal{O}}, ||\mathcal{C}_{\mathcal{O}} - P_2||_2 = \mathcal{R}_{\mathcal{O}} \end{aligned} \tag{26}$$

Then, we can define

$$arcsin \frac{\mathcal{R}_{\mathcal{O}}}{||\mathcal{C}_{\mathcal{O}} - \nu||_2} \leq \theta_{\mathcal{O}} \leq 2\pi - arcsin \frac{\mathcal{R}_{\mathcal{O}}}{||\mathcal{C}_{\mathcal{O}} - \nu||_2} \tag{27}$$

Similarly, we obtain the angle range of $\theta_{\mathcal{I}}$

$$arcsin \frac{\mathcal{R}_{\mathcal{I}}}{||\mathcal{C}_{\mathcal{I}} - \nu||_2} \leq \theta_{\mathcal{I}} \leq 2\pi - arcsin \frac{\mathcal{R}_{\mathcal{I}}}{||\mathcal{C}_{\mathcal{I}} - \nu||_2} \tag{28}$$

Because $\mathcal{R}_{\mathcal{O}} = (1 + \epsilon)\mathcal{R}_{\mathcal{I}}$, we then have

$$arcsin \frac{\mathcal{R}_{\mathcal{I}}}{||\mathcal{C}_{\mathcal{I}} - \nu||_2} < arcsin \frac{\mathcal{R}_{\mathcal{O}}}{||\mathcal{C}_{\mathcal{O}} - \nu||_2} \tag{29}$$

So, the lemma follows. ■

Lemma 11. *For crossed MEBs, training the data of outer volume may lead to a very high error rate in the hyperplane fitting. Suppose that the data are evenly distributed in the MEB, the error rate of classification on this pair of MEBs is at most $\Omega(\frac{1}{(1+\epsilon)^m})$.*

Proof. By Remark 2, we know the parameter ϵ decides the radius of $B(\mathcal{R}_{\mathcal{O}}, \mathcal{C}_{\mathcal{O}})$. Assume the data space is close to an normal distribution, we can find

$$\frac{Vol(\mathcal{O}) - Vol(\mathcal{I})}{Vol(\mathcal{O})} \leq \frac{Vol(\mathcal{O}) - \frac{Vol(\mathcal{O})}{(1+\epsilon)^m}}{\mathcal{O}} = \frac{1}{(1+\epsilon)^m} \tag{30}$$

So, the lemma follows. Because training the data of outer volume may lead to a high error rate, the classification hypothesis has a high probability to be a null hypothesis. Then, Theorem 9 follows. ■

5.2. Finding the optimal inscribed hypersphere

To represent the version space, we use MMD (Maximum Mean Discrepancy) as the metric function which can measure the difference of two distributions. The kernel type of it is described as follows:

$$\begin{aligned} &MMD^2(\mathcal{F}, X, Y) := \\ &\frac{1}{\mathbf{m}^2}\sum_{i=1}^{\mathbf{m}} k(x_i, x_j) - \frac{2}{\mathbf{mn}}\sum_{i,j=1}^{\mathbf{mn}} k(x_i, y_i) + \frac{1}{\mathbf{n}^2}\sum_{i,j=1}^{\mathbf{n}} k(y_i, y_j) \end{aligned} \tag{31}$$

where $k(\cdot, \cdot)$ denotes the kernel metric of the two input objects. Suppose that the kernel function is bounded, i.e., $k(\cdot, \cdot) \leq \kappa$, we have the following upper bound and lower bounds of kernel MMD,

$$\begin{aligned} 0 \leq &MMD^2(\mathcal{F}, X, Y) \leq (\mathbf{m} + \mathbf{n})\kappa - \frac{2}{\mathbf{mn}}\sum_{i,j=1}^{\mathbf{mn}} k(x_i, y_i) \\ &s.t.\ MMD^2(\mathcal{F}, X, Y) = 0,\ iff\ \mathbf{m} = \mathbf{n} \end{aligned} \tag{32}$$

In AL sampling, the optimization objective is to minimize the original space $\mathcal{H}$ and representation space $\mathcal{D}'$,

$$\min_{\mathcal{D}' \subset \mathcal{D}} MMD^2(\mathcal{D}, \mathcal{D}') \tag{33}$$

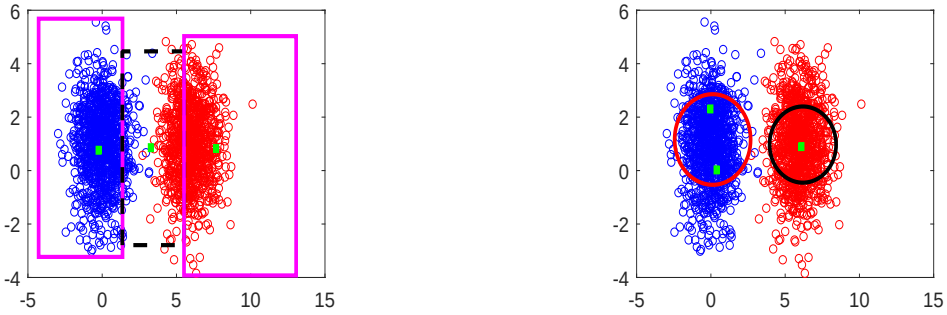
Therefore, we need to minimize the upper bound of Eq. (32), that is to say,

$$\min_{h_i \in \mathcal{D}', h_j \in \mathcal{D}} \frac{2}{\mathbf{mn}}\sum_{i,j=1}^{\mathbf{mn}} k(h_i, h_j) \tag{34}$$

To minimize it, we need to optimize the local representation sampling process via associating y_i within the local space of x_i. Assume the querying number is $\mathcal{K}$, the structure loss of representative space can be defined as:

$$min\ \left\{ \mathcal{L}(\mathcal{D}', \mathcal{D}) = \sum_{i}^{\mathcal{K}} MMD(\mathcal{V}_i, B(\mathcal{C}_{\mathcal{V}_i}, \mathcal{R}_{\mathcal{V}_i})) \right\} \tag{35}$$

where $\mathcal{D}' = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_\mathcal{K}, \}$ and $B(\mathcal{C}_{\mathcal{V}_i}, \mathcal{R}_{\mathcal{V}_i}))$ is its local inscribed hypersphere of $\mathcal{V}_i$.



(a) Representation by inner volume

(b) Representation by outer and inner volumes

Figure 3: (a)This illustration is an example of noise bias of AL by inner volume, where the three green points are the representation hypothesis of the original version space, and the there rectangle areas are their local spaces. Intuitively, the noises will make the classifier be hard to separate the two classes and then lead to a high error rate AL result. (b) This illustration shows the representation sampling result by inner and outer volumes, where the two circles represent the outer volume of the original version space, and the three green points are the representation hypothesis within the enclosing space of outer volume. By observation, we could find this learning way smooth the noises bias and can provide a better AL sampling guidance since it removes the outliers before finding representation samples.

5.3. Excluding the outlier hypotheses

However, there exists a sampling bias (see Figure 3 in Euclidean space). To describe this kind of situation, here we highlight it in the following remark.

Remark 5. *An outlier hypothesis h_Φ may lead to a fast local convergence when $\mathcal{C}_{\mathcal{V}_i} = h_\Phi, \exists \mathcal{V}_i$.*

To exclude the outlier hypotheses, we need to remove the hypotheses distributed outside the outer volume. By *Lemma 10*, we mark the hypotheses located outside the inner volume be outlier hypotheses, that is to say

Remark 6. *For arbitrary hypothesis in the $B(\mathcal{C}_I, \mathcal{R}_I)$, the hypothesis which satisfies Eq.(28) is a outlier hypothesis.*

Considering theses outlier hypotheses having low relevance to its local neighbor hypothesis, we observe the volume of the local inscribed hypersphere is bigger than the non-outlier hypotheses. Therefore, we propose the ϵ' approximation split approach to define the outlier hypothesis:

$$\frac{Vol(B(h_\Phi))}{Vol(\mathcal{O})} > \epsilon' \tag{36}$$

where $B(h_\Phi)$ represents the MEB that covers h_Φ.

After the above discussion, we present our theoretical AL proposition.

Proposition 3. *Let $\mathcal{O}'$ represent the enclosing space of $\mathcal{O}$ which has removed all outlier hypotheses, here we present our AL sampling objective function:*

$$min \left\{ \mathop{\mathcal{L}(\mathcal{D}', \mathcal{O}')}_{B \subset \mathcal{D}' \subset \mathcal{O}'} = \sum_{i}^{\mathcal{K}} MMD(\mathcal{V}_i, B(\mathcal{C}_{\mathcal{V}_i}, \mathcal{R}_{\mathcal{V}_i})) \right\} \tag{37}$$

5.4. Proposed active learning model

In this section, we present our theoretical AL model in algorithm 1. To exclude the outlier hypotheses, Step 4 to 9 remove the hypotheses that located outside the outer volume. Then, we find a local optimal representation for the current version space via an EM learning process in Step 11 to 16. Finally, we return the centers of each hypersphere as AL sampling examples.

Algorithm 1: Volum-based Active Learning Model

1 **Input:** Version space $\mathcal{H}$
2 Volume-splitting parameter: ϵ'
3 **Begin:**
4 **for** *1* $\leftarrow \Phi$ *to* k **do**
5 **if** $\frac{Vol(B(h_\Phi))}{Vol(\mathcal{O})} > \epsilon'$ **then**
6 Remove h_Φ from $\mathcal{H}$.
7 Update $\mathcal{H}$.
8 **end**
9 **end**
10 Obtain the enclosing space of $\mathcal{O}$: $\mathcal{O}' \leftarrow \mathcal{H}$.
11 Initialize $\mathcal{D}'$ by passive querying: $\mathcal{D}' = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_\mathcal{K}\}$.
12 **while** $\mathcal{L} - \mathcal{L}' \neq 0$ **do**
13 Calculate the loss function $\mathcal{L}(\mathcal{D}', \mathcal{O}') = \sum_i^\mathcal{K} MMD(\mathcal{V}_i, B(\mathcal{C}_{\mathcal{V}_i}, \mathcal{R}_{\mathcal{V}_i}))$.
14 Update their MEBs of $\{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_\mathcal{K}\}$.
15 Update the loss function $\mathcal{L}'$.
16 **end**
17 Return the centers of the final MEBs.

6. Expectation Maximization in Sparse Space

Although half-space can reduce the volume of the version space, cutting which half-space is hard to decide whatever in the version space or Euclidean space since the number of half-space is infinite. However, shrinking the volume of the version space is effective. In this section, we propose a new shrinking method by reducing the number of candidate hypothesis set. In non-line feature space, spanned by kernel, we use sequential optimization to map the original kernel space into a spare space by halving the size of kernel space. Compared with half-space, the sparse space has two advantages: exclude outlier hypotheses, and remove the similar hypotheses of arbitrary local spaces. For the sparse space, it optimizes a global representation in the enclosing space of outer volume of the data space. Then, we find that the EM model which returns the local centers can have an effective local representation optimization.

In this section, Section 6.1 describes the global sparse space by halving the size of input space, Section 6.2 discusses the effectiveness of EM model which returns the local centers for representation sampling, Section 6.3 describes the Volume-based AL algorithm, and Section 6.4 discusses the time and space complexities of VAL.

6.1. Global sparse by halving

In machine learning community there have been extensive experimental design approaches. Among them, transductive experimental design is one effective optimization scheme which acts on active learning issues.

Considering a linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ from measurements [40] $y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i$, where $w \in \mathcal{R}^d$, and $\epsilon_i\,\mathcal{N}(0, \sigma^2)$. AL sampling is to optimize a set of $\mathbf{V} = \{(v_1, y_1), (v_2, y_2), ..., (v_m, y_m)\}$ to represent $\mathbf{x}$. Therefore, the MLE (maximum likelihood estimate) of $\mathbf{w}$ is obtained by

$$\underset{\mathbf{w}^*}{argmin}\left\{\mathcal{J}(\mathbf{w}) = \sum_{i=1}^{n}(\mathbf{w}^T z_i - y_i)\right\} \tag{38}$$

and the error rate is

$$\begin{aligned} e &= w - \mathbf{w}^* \\ s.t.\ \mu(e) &= 0, D(e) = \sigma^2\mathbf{C}_w \end{aligned} \tag{39}$$

where $\mu(\cdot)$ denotes the mean value of the input variable, $D(\cdot)$ denotes the covariance matrix of the input object, and

$$\mathbf{C}_w = \left(\frac{\partial^2\mathcal{J}}{\partial\mathbf{w}\mathbf{w}^T}\right)^{-1} = (\mathbf{V}\mathbf{V}^T)^{-1} \tag{40}$$

Then the average expected square predictive error over $\mathcal{X}$ can be wrote as

$$E(y_i - w^* T x_i) = \sigma^2 + \sigma^2 \mathbf{Tr}(\mathcal{X}^T \mathbf{V}\mathbf{V}^T \mathcal{X}) \tag{41}$$

Therefore, the optimization objective function is:

$$\begin{aligned} \underset{\mathbf{V},\mathbf{A}}{argmin} \sum_{1}^{n} ||\mathbf{x}_i - \mathbf{V}^T \alpha_i|| + \mu ||\alpha_i|| \\ \mathbf{V} \subset \mathcal{X}, \mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_n] \end{aligned} \tag{42}$$

After mapping the original input space into an non-linear kernel space, we iteratively project the top-($\lfloor n/2 \rfloor$) data with high confidence scores to a sparse space by sequential optimization, where the confidence score of the optimization is described as follows:

$$\begin{aligned} \mathcal{C}(x_i) &= \frac{||\mathbf{K}(l,:)\mathbf{K}(:,l)||^2}{\mathbf{K}(l,l) + \mu}, \forall i \\ s.t.\ \mathbf{K} &\leftarrow \mathbf{K} - \frac{\mathbf{K}(:,l')\mathbf{K}(l',:)}{\mathbf{K}(l',l') + \mu} \end{aligned} \tag{43}$$

where $\mathbf{K}$ is the kernel matrix of $\mathcal{X}$, l represents the sequence position of x_i in $\mathcal{X}$, and l' represents the sequence position of the data with current highest confidence score in $\mathcal{X}$.

6.2. Center representation by EM

Interestingly, the above optimization is a global optimization scheme which satisfies:

$$\underset{\mathbf{V}}{argmin}\ MMD(\mathbf{V}, \mathcal{X}) \tag{44}$$

and it can not guarantee a local optimization solution which satisfies:

$$\underset{\mathbf{V}}{argmin}\ MMD(\mathcal{V}, \mathcal{X}) + \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} MMD(v_i, \mathcal{S}_i) \tag{45}$$

where $\mathcal{S}_i$ is the represented local space of $\mathcal{S}_i$.

Considering the MMD metric learning, we observe that

Theorem 12. *Center representation can meet the optimization requirement of Eq. (39).*

Proof. Let

$$v_i = \sum_{x_i \in \mathcal{S}_i} x_i \tag{46}$$

By this setting, we have the following results: $MMD(v_i, \mathcal{S}_i) = 0, \forall i$, and $\frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} MMD(v_i, \mathcal{S}_i) = 0$, then, $MMD(\mathbf{V}, \mathcal{X}) = 0$. Therefore, Eq. (45) will be zero and it is the lower bound. ■

Algorithm 2: Volume-based AL

Input:

$\mathcal{X}$: data set with size of $n \times m$.

$\mathcal{K}$: number of queries.

Initialize:

$\mathbf{K}$: the kernel matrix of $\mathcal{X}$.

Sparse matrix $\mathcal{X}'$: $\leftarrow \mathcal{X}^*$.

$1 \leftarrow l, l', j, t$.

Begin:

Calculate the kernel matrix $\mathbf{K}$ of $\mathcal{X}$.

while $t \leq \lfloor n/2 \rfloor$ **do**

  for *each data point* $x_i \in \mathcal{X}$ **do**

    Calculate the confidence score of x_i: $\mathcal{C}(x_i) = \frac{||\mathbf{K}(l,:)\mathbf{K}(:,l)||^2}{\mathbf{K}(l,l)+\mu)}$.

    Select the data point with the highest confidence score and add it to matrix $\mathcal{X}^*$.

    Update $\mathbf{K}$ by $\mathbf{K} \leftarrow \mathbf{K} - \frac{\mathbf{K}(:,l')\mathbf{K}(l',:)}{\mathbf{K}(l',l')+\mu}$.

    Update t: $t \leftarrow t + 1$.

  end

end

Initialize $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_\mathcal{K}\}$ by passive sampling in $\mathcal{X}^*$.

while *j* **do**

  Divide the local space to $\mathcal{K}$ parts by the model Θ: $B = \{B_1, B_2, ..., B_\mathcal{K}\}$.

  Update $\mathcal{U}' = \{\hat{u}_1, \hat{u}_2, ..., \hat{u}_\mathcal{K}\}$.

  Calculate the loss functions of $\mathcal{U}$ and $\mathcal{U}'$ by $\mathcal{L}_j = \sum_{i=1}^{\mathcal{K}} \ell(\mathcal{C}_i, B_i)$ and $\mathcal{L}_{j+1} = \sum_{i=1}^{\mathcal{K}} \ell(\hat{u}_i, B_i)$.

  if $\mathcal{L}_j - \mathcal{L}_{j+1} \to 0$ **then**

    Break

  end

  Update j: $j \leftarrow j + 1$.

end

Query the labels of $\mathcal{U}$ and store them in matrix y.

Train the classification model h on $(\mathcal{U}, y)$.

Predict $\mathcal{X}$ on h.

Return error rate on $\mathcal{X}$.

In order to minimize Eq. (45), here we propose the objective function of our representative learning approach:

$$\underset{\Theta,\ell}{argmin} \left\{ \mathcal{L} = \sum_{i}^{\mathcal{K}} \Theta(\mathcal{X}^*, \ell, k)) \right\} \tag{47}$$

where Θ represent the metric model of local space division, ℓ is the local measurement of data points, and $\mathcal{K}$ is the querying number of AL.

6.3. Proposed VAL algorithm

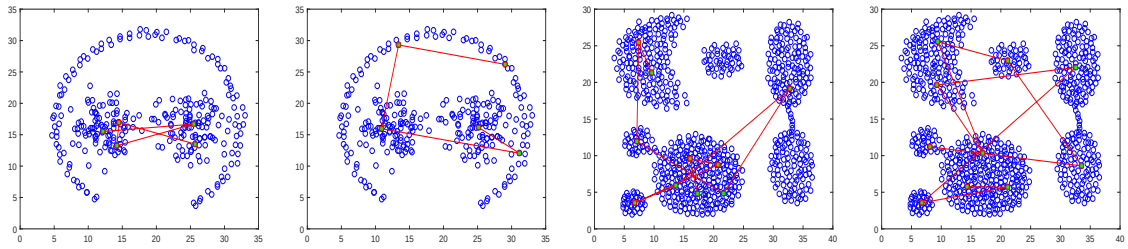
Based on the above model definitions and analysis, we design an executable algorithm in this section, called VAL. To remove the non-promising hypotheses and have a global representation for original data space, Step 10 to 17 use the sequential optimization to reduce the number of data set into a half, in which the method favours to select the data with the highest confidence score in the current kernel matrix.

After obtaining the sparse space $\mathcal{X}^*$, Step 18 to 27 use the EM iteration to minimize the local representation loss. In the iteration process, we define the model Θ as $||.,.||_2^2$ to classify the current data sets to $\mathcal{K}$ local areas, and we define the metric function of the loss function as $\ell = ||.,.||_2^2$. After the convergence, we use the centers of final local segmentation areas for AL sampling. Finally, Step 28 queries the labels of representation set $\mathcal{U}$, Step 29 trains a classifier on it, and then Step 30 to 31 return the prediction error rate on $\mathcal{X}$.

6.4. Time and space complexities

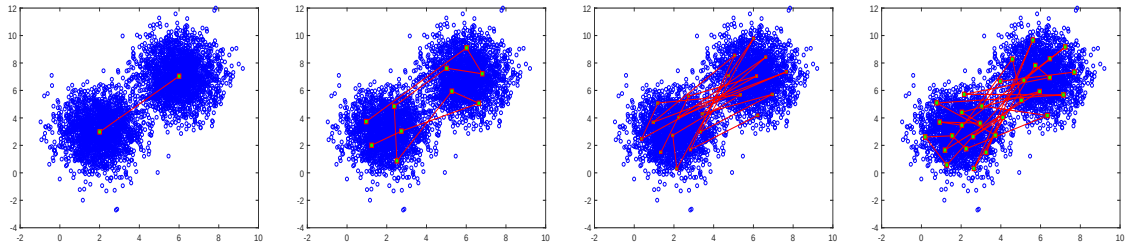
The general AL strategies, which use the prior labeled set to guide the unseen process of sampling, depends heavily on the size of initialization input. Then, the time cost of outputting the label space is decided by classifier and input set. For example, let $\mathcal{T}$ be the sample number of input space, and SVM be the classifier, then the time cost of one training will be $O(\mathcal{T}^2)$ to $O(\mathcal{T}^3)$. To select $\mathcal{K}$ samples, the time cost can be loosely described as $O(\mathcal{K}\mathcal{T}^2)$ to $O(\mathcal{K}\mathcal{T}^3)$. Moreover, the space cost of SVM is $O(\mathcal{T}^2)$ to $O(\mathcal{T}^3)$, and this consumption might be the minimum space cost of the AL. Therefore, the time and space complexities of the AL strategies which depend on training model and labeled set are "uncertain".

However, our VAL algorithm is a target-independence approach which does not depend on the labeled set and classifiers, and its time and space complexities are "certain". In its two main steps, the kernel matrix costs a time complexity of $O(n^2)$, and the EM model approximately costs a time complexity of $O(n^2)$. In space consumption, the space price is about $O(n^2)$.



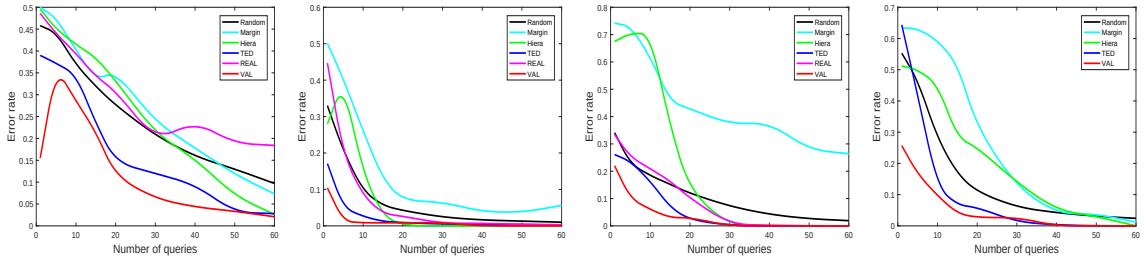
(a) TED on Pathbased (b) VAL on Pathbased (c) TED on Aggregation (d) VAL on Aggregation

Figure 4: Representation structure of Pathbased and Aggregation data sets by TED and VAL. The green points are the representation data points, and the line represents the space structure. The observation shows our proposed VAL algorithm has a better space representation than that of TED since TED only considers the global representation, while VAL uses global and local optimizations for the representation.

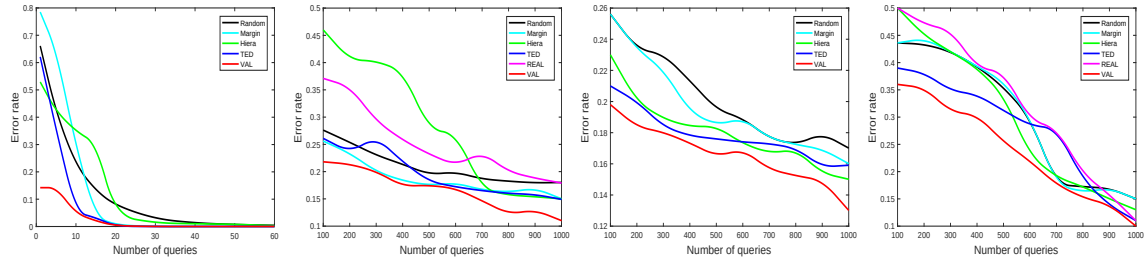


(a) $\mathcal{K} = 2$ (b) $\mathcal{K} = 10$ (c) $\mathcal{K} = 20$ (d) $\mathcal{K} = 30$

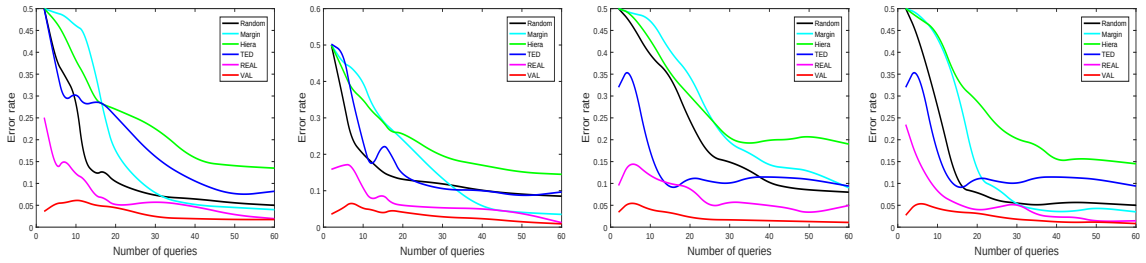
Figure 5: An example of representation process of VAL, where $\mathcal{K}$ is the sampling number. We can observe that the representation results are very effective no matter how many sampling numbers we set since VAL optimizes the representation process globally and locally.



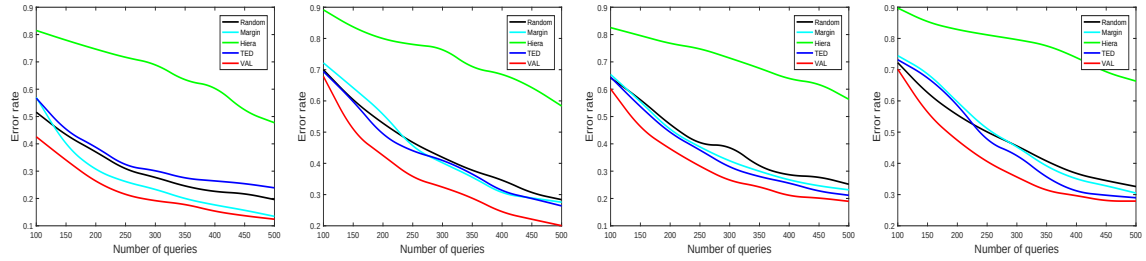
(a) g2-2-30: 2,048×2×2 (b) Flame: 240×2×2 (c) Jain: 373×2×2 (d) Pathbased: 300×2×3



(e) Aggregation: 788×2×7 (f) Adult: 1605×123×2 (g) Satimage: 4,435×36×6 (h) Phishing:11,055×68×2



(i) DvsP: 1,608×16×2 (j) EvsF: 1,543×16×2 (k) MvsN: 1,575×16×2 (l) UvsV: 1,577×16×2



(m) A-P: 3,096×16×4 (n) A-T: 9,148×12×2 (o) A-X: 15,364×20×2 (p) A-Z: 20,000×16×26

Figure 6: The error rate curves of different AL approaches on structured and unstructured data sets by training LIBSVM. (a)-(e) are the clustering data sets. (f)-(h) are three UCI real-world data sets. (i)-(l) are the selected binary classification sub data sets of *letter*. (n)-(t) are the selected multi-class sub sets of *letter*, where A-P, A-X, A-T, A-Z represents the letter sub sets of A to P, X, Z, respectively. For fair comparison, we select one data point from each class to run the compared AL algorithms in all the experimental data sets, respectively. These initialization points also are added to the training set of VAL. The *Number of queries* of all the figures represent the selected data points in AL sampling.

7. Experiments

Because the proposed VAL algorithm is based on structured version space theory, this section will report the comparison experiments on some structured data sets (classical clustering data sets) and observe its performance in an unstructured data set (letter image recognition data set *letter*). Related baseline approaches which compare VAL are introduced as follows:

- Random: takes the idea of random sampling and can be adapted in any data setting, but not stable.
- Margin: selects the data point with the closet distance to the current classification model from the pool in each iterative sampling. It is a classical AL algorithm based on SVM.
- Hierarchical: judgments the cluster subtree whether can be labeled with the root node's label based on a probability function. It is clustering-based AL, and it connects the unsupervised learning in AL.
- TED: prefers the data points that are on the one side hard to-predict and on the other side representative for the rest of the pool. It is Transductive Experimental Design work in statistics AL. Similar woks can been seen in Optimum Experimental Design (OED), D, A, and E-optimal Design.
- Re-active: selects the data points which have the biggest influence on current prediction model after querying. It maximizes the model differences to sample. Whatever kinds of classifiers could be trained in the relabelling learning.

In additional, error rate is used to evaluate the classification result in this paper and the lowest classification error rates of each algorithm are reported in Figure 6, where LIBSVM [50] is the trained classifier. Before the experiments, we give two examples for our representation approach.

7.1. Examples of representation results

TED is a good representative learning approach for global optimization. Our VAL is based on global and local optimization. The difference between the two algorithms is whether the whole geometric structure of the data is represented and mapped. To get a good visual result of how they perform differently, Figure 4 shows two examples on two clustering data sets. As seen, TED loses the representative structure in classes with weak clustering features, but our algorithm has a better represent since it further considers the local optimization. Figure 5 also reports a group sampling process by finding the optimal representation of VAL. It will help us to understand the sampling process of our proposed AL approach.

7.2. Performance on structured data sets

Eight classical two-dimension clustering data sets are tested in this section. They are challenging clustering tasks with one or more characteristics of adjacent classes, a lot of noises, linear inseparable, multi-density, etc. To run the approaches which need the support of labeled data, annotating one data point from each class is our special data preprocessing work that aims to provide various label information. Otherwise, missing one or more kinds of labels will lead to a biased supervised learning. Therefore, avoiding the negative influence of label kinds in our data preprocessing method is necessary for optimizing Margin, Hierarchical, Re-active.

In the reported results of Figure 6(a)-(h), Random is stable but not prominent under a random sampling strategy through observing that its error rate curve is located in the middle position of the six curves. Margin is easy to be influenced by the noises located near the classification model with fuzzy labels according the bad performance in this group of experiment. Hierarchical clustering provides other prior knowledge of class structure for the future probability model of active annotating. However, this approach depends on the clustering results, and the error rate will increase quickly if the precision rate of clustering is low. TED has a stable representative sampling strategy and shows low error rates in this group experiment. But the sensitivity of parameters setting is higher than others. Re-active observes the model parameter change when annotating the unlabeled data in positive or negative label and then selects the data which can maximize this difference. While noises always misled their choice because they may change the training model seriously with a fuzzy class label. Therefore, it performs not well in the clustering data sets with a lot of noises. For VAL, it performs best in the seven clustering data sets, compared with others, since the representation space has a high effective representation of original space after removing all outliers.

7.3. Performance on unstructured data sets

This data set is to identify the 26 capital letters in the type of black-and-white rectangular pixel with the total number of 20,000 images that are converted into 20,000$\times$16 numerical matrix and each element is scaled to fit into a range of integer values from 0 through 15. But it does not have clear cluster structures and we have used different unsupervised clustering approaches to test them. Before this group of experiments, we select 5 and 7 groups letters as binary classification, multi-class tasks, respectively. The curves of each baselines' lowest classification error rate on this group of experiment have been drawn in Figure 6(i)-(l).

By observing the error rate curves of the seven different approaches in the two-classification experiments, their difference increases clearly in the high dimensional space, compared with the low dimensional experiments of previous section. In the drawn curves, Random and Margin still keep their characteristics, and their performance are similar with the last group of experiment. But Hierarchical performs badly since there are no clear

cluster structures. Lack of the correct guidance of clustering results, its probability evaluation model is unstable and then leads to a wrong active annotating result. For TED, the error rate of it also begins to raise in this unstructured data space. But in the results of Re-active and VAL, their error rates decline rapidly in these noiseless data sets, but the advantage of the latter is more outstanding.

The change of the drawn curves in Figure 6(m)-(p) shows the classification results of different baselines in the multi-class setting, where our proposed VAL algorithm shows significant advantage since the local optimization still works in the unstructured data set. This group of experiment evaluates that our proposed VAL algorithm can reduce the error rate rapidly with low querying cost whatever in two or multi-classification problem.

8. Discussions

In this section, we present two-fold discussions on our proposed perspective of volume sampling.

8.1. The gap between volume and supervision

Volume sampling is a highly abstract process of finding the ϵ-optimal hypothesis. Based on a structured perspective, we divide the volume of version space into two parts: outer and inner volumes, where the hypotheses distributed in the outer volume connect a kind of informative samples in real-world unlabeled data, and the hypotheses of inner volume map the most of representative unlabeled data distributed in the internal regions of the clusters.

In real-world scenarios, the outer volume sampling strategies need the supervision of classifier and labeled set. While the inner volume sampling can use an unsupervised way to find a group of representative samples or a supervised policy to minimize the distribution distance between the labeled and unlabeled sets. Therefore, outer volume sampling cannot break the curse of supervision and inner volume sampling can be dependent of supervision. For our proposed VAL algorithm, it does not need the supervision of classifier and labeled set. In addition, outer volume sampling strategies need to cooperate the diversity [51] of the labels in multi-class sampling tasks. However, inner volume sampling will not consider such a special label setting since it only focuses on the distribution of instances.

8.2. The application scenarios of volume sampling

AL sampling has two types of query scenarios including single query and batch query [52], where single query returns one sample for the learner at one time, and bath query requires the algorithm to return a group of unlabeled data at each query. Generally, single query approaches pay attention on the query principle, while bath query considers the sampling tasks of large-scale data sets [53]. It is thus, the used classifiers in single

query are fast and cheap in terms of time or space prices such as SVM. After producing robust single query strategies, learners consider to use more complex classifiers to query a group of unlabeled data for annotation such as Bayesian network [54], deep learning frameworks [55].

In this study, our proposed VAL strategy aims to provide theoretical guarantees for cooperating two different kinds of outer and inner volume sampling ideas in single query task setting. The goal is to accelerate the decline of the error rate of prediction using as few labels as possible. From a geometrical view, the outer volume sampling approach can be adopted in the detection of outliers and cluster boundary points [56]. Without a fixed distribution assumption, it can provide model guidance for the geometrical sampling [57] in different task settings such as adversarial examples [58]. For the inner volume sampling, it can benefit the representation learning, autoencoder, and etc.

9. Conclusion

Lack of enough label support motivated different types of AL sampling strategies to query more labels of unlabeled data to improve the training, such as iterative sampling by uncertainty evaluation and maximization of model hypothesis. However, available algorithms are in a supervised way which requires enough label information in terms of a task specific setting. To reduce the target-dependence of labeled set, it motivates us to consider which sampled data can maximize the classification hypothesis or distribution update in the version space after adding them to training set.

In this paper, we study the outer and inner volumes of version space, where the hypothesis set of outer volumes could maximize the hypothesis distance between current and updated classification hypothesis, and the hypothesis set of inner volume represents the learned representation structure of data distribution. While neither outer volume or inner volume can produce a highly representation to version space, we find the optimal representation of inner volume in the enclosing space of outer volume, and further proposed the VAL algorithm. Experimental results of the proposed algorithm have shown that it can reach the optimal prediction rapidly with a few number of queries and the decline rate of error rate is faster than the other compared approaches. In future work, we will further study the relationship of outer and inner volumes in the version space.

Acknowledgement

This paper was independently finished by Xiaofeng Cao when he was a PhD candidate with Advanced Analytics Institute, University of Technology Sydney (09.2018-05.2019). Thanks for the financial support of

Pro.Guandong Xu during this period. He now is pursuing the Ph.D. degree with the Centre for Artificial Intelligence, University of Technology Sydney (06.2019-). Thanks for the supervision of Pro. Ivor Tsang on this study.

References

[1] C. Campbell, Y. Ying, Synthesis lectures on artificial intelligence and machine learning, Learning with support vector machines 5 (2011) 1–95.

[2] S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017.
URL http://www.aaai.org/Library/AAAI/aaai17contents.php

[3] L. Zhang, H. P. H. Shum, L. Shao, Manifold regularized experimental design for active learning, IEEE Trans. Image Processing 26 (2) (2017) 969–981. doi:10.1109/TIP.2016.2635440.
URL https://doi.org/10.1109/TIP.2016.2635440

[4] B. Du, Z. Wang, L. Zhang, L. Zhang, D. Tao, Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion, IEEE Transactions on Image Processing 26 (4) (2017) 1694–1707.

[5] S. Huang, J. Chen, X. Mu, Z. Zhou, Cost-effective active learning from diverse labelers, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, 2017, pp. 1879–1885.

[6] A. Gonen, S. Sabato, S. Shalev-Shwartz, Efficient active learning of halfspaces: an aggressive approach, The Journal of Machine Learning Research 14 (1) (2013) 2583–2615.

[7] T. M. Mitchell, Generalization as search, Artif. Intell. 18 (2) (1982) 203–226. doi:10.1016/0004-3702(82)90040-6.
URL https://doi.org/10.1016/0004-3702(82)90040-6

[8] Y.-F. Li, I. W. Tsang, J. Kwok, Z.-H. Zhou, Tighter and convex maximum margin clustering, in: Artificial Intelligence and Statistics, 2009, pp. 344–351.

[9] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of machine learning research 2 (Nov) (2001) 45–66.

[10] S. Dasgupta, Coarse sample complexity bounds for active learning, in: Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], 2005, pp. 235–242.

[11] J. Kremer, K. S. Pedersen, C. Igel, Active learning with support vector machines, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4 (4) (2014) 313–326. doi:10.1002/widm.1132.
URL https://doi.org/10.1002/widm.1132

[12] C. Tosh, S. Dasgupta, Diameter-based active learning, ICML.

[13] S. Dasgupta, D. J. Hsu, C. Monteleoni, A general agnostic active learning algorithm, in: International Symposium on Artificial Intelligence and Mathematics, ISAIM 2008, Fort Lauderdale, Florida, USA, January 2-4, 2008, 2008.

[14] A. Gonen, S. Sabato, S. Shalev-Shwartz, Efficient active learning of halfspaces: an aggressive approach, Journal of Machine Learning Research 14 (1) (2013) 2583–2615.
URL http://dl.acm.org/citation.cfm?id=2567744

[15] M. Balcan, A. Beygelzimer, J. Langford, Agnostic active learning, in: Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, 2006, pp. 65–72.

[16] D. A. Cohn, L. E. Atlas, R. E. Ladner, Improving generalization with active learning, Machine Learning 15 (2) (1994) 201–221. doi:10.1007/BF00993277.
URL https://doi.org/10.1007/BF00993277

[17] G. R. Brightwell, P. Winkler, Counting linear extensions is #p-complete, in: Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, May 5-8, 1991, New Orleans, Louisiana, USA, 1991, pp. 175–181.

[18] T. M. Chan, Approximating the diameter, width, smallest enclosing cylinder, and minimum-width annulus, International Journal of Computational Geometry & Applications 12 (01n02) (2008) 67–85.

[19] M. Badoiu, K. L. Clarkson, Optimal core-sets for balls, Comput. Geom. 40 (1) (2008) 14–22. doi:10.1016/j.comgeo.2007.04.002.
URL https://doi.org/10.1016/j.comgeo.2007.04.002

[20] P. Kumar, J. S. B. Mitchell, E. A. Yildirim, Approximate minimum enclosing balls in high dimensions using core-sets, ACM Journal of Experimental Algorithmics 8. doi:10.1145/996546.996548.
URL https://doi.org/10.1145/996546.996548

[21] I. W. Tsang, J. T. Kwok, P.-M. Cheung, Core vector machines: Fast svm training on very large data sets, Journal of Machine Learning Research 6 (Apr) (2005) 363–392.

[22] I. W. Tsang, J. T. Kwok, J. M. Zurada, Generalized core vector machines, IEEE Trans. Neural Networks 17 (5) (2006) 1126–1140. doi:10.1109/TNN.2006.878123.
URL https://doi.org/10.1109/TNN.2006.878123

[23] I. W. Tsang, J. T. Kwok, Large-scale sparsified manifold regularization, in: Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006, 2006, pp. 1401–1408.

[24] I. W. Tsang, A. Kocsor, J. T. Kwok, Simpler core vector machines with enclosing balls, in: Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007, 2007, pp. 911–918.

[25] M. Badoiu, S. Har-Peled, P. Indyk, Approximate clustering via core-sets, Proc.annu.acm Sympos.theory Comput (2002) 250 – 257.

[26] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167. doi:10.1023/A:1009715923555.
URL https://doi.org/10.1023/A:1009715923555

[27] P. M. Hubbard, Approximating polyhedra with spheres for time-critical collision detection, ACM Trans. Graph. 15 (3) (1996) 179–210. doi:10.1145/231731.231732.
URL https://doi.org/10.1145/231731.231732

[28] N. V. Cuong, N. Ye, W. S. Lee, Robustness of bayesian pool-based active learning against prior misspecification, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[29] Y. Guo, G. Ding, Y. Gao, J. Wang, Semi-supervised active learning with cross-class sample transfer., in: IJCAI, 2016, pp. 1526–1532.

[30] P. Bachman, A. Sordoni, A. Trischler, Learning algorithms for active learning (2017) 301–310.

[31] S. Dasgupta, D. Hsu, Hierarchical sampling for active learning, in: ICML, 2008, pp. 208–215.

[32] W. Hu, W. Hu, N. Xie, S. Maybank, Unsupervised active learning based on hierarchical graph-theoretic clustering, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (5) (2009) 1147–1161.

[33] H. T. Nguyen, A. W. M. Smeulders, Active learning using pre-clustering, in: Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, 2004.

[34] M. Wang, F. Min, Z.-H. Zhang, Y.-X. Wu, Active learning through density clustering, Expert Systems with Applications 85 (2017) 305–317.

[35] S.-J. Huang, R. Jin, Z.-H. Zhou, Active learning by querying informative and representative examples, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (36) (2014) 1936–1949.

[36] C. H. Lin, M. Mausam, D. S. Weld, Re-active learning: Active learning with relabeling., in: AAAI, 2016, pp. 1845–1852.

[37] A. K. McCallumzy, K. Nigamy, Employing em and pool-based active learning for text classification, in: ICML, Citeseer, 1998, pp. 359–367.

[38] G. Schohn, D. Cohn, Less is more: Active learning with support vector machines, in: Seventeenth International Conference on Machine Learning, 2000.

[39] X. Zhao, Y. Kai, V. Tresp, X. Xu, J. Wang, Representative sampling for text classification using support vector machines, in: European Conference on Ir Research, 2003.

[40] K. Yu, J. Bi, V. Tresp, Active learning via transductive experimental design, in: ICML, 2006, pp. 1081–1088.

[41] W. Zheng, J. Ye, Querying discriminative and representative samples for batch mode active learning, Acm Transactions on Knowledge Discovery from Data 9 (3) (2015) 1–23.

[42] Z. Lijun, C. Chun, B. Jiajun, C. Deng, H. Xiaofei, T. S. Huang, Active learning based on locally linear reconstruction, IEEE Transactions on Pattern Analysis & Machine Intelligence 33 (10) (2011) 2026–2038.

[43] S. Dasgupta, Analysis of a greedy active learning strategy, in: Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada], 2004, pp. 337–344.

[44] S. Dasgupta, Coarse sample complexity bounds for active learning, in: NIPS, 2006, pp. 235–242.

[45] Y. Guo, G. Ding, Y. Gao, J. Han, Active learning with cross-class similarity transfer, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[46] L. Shi, Y.-D. Shen, Diversifying convex transductive experimental design for active learning., in: IJCAI, 2016, pp. 1997–2003.

[47] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, D. Tao, Exploring representativeness and informativeness for active learning, IEEE transactions on cybernetics 47 (1) (2017) 14–26.

[48] R. Wang, X.-Z. Wang, S. Kwong, C. Xu, Incorporating diversity and informativeness in multiple-instance active learning, IEEE Transactions on Fuzzy Systems 25 (6) (2017) 1460–1475.

[49] J. B. Lasserre, A generalization of löwner-john's ellipsoid theorem, Math. Program. 152 (1-2) (2015) 559–591. doi:10.1007/s10107-014-0798-5.
URL https://doi.org/10.1007/s10107-014-0798-5

[50] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 27.

[51] Y. Yang, Z. Ma, F. Nie, X. Chang, A. G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, International Journal of Computer Vision 113 (2) (2015) 113–127.

[52] Y. Guo, D. Schuurmans, Discriminative batch mode active learning, in: Advances in neural information processing systems, 2008, pp. 593–600.

[53] S. C. Hoi, R. Jin, M. R. Lyu, Large-scale text categorization by batch mode active learning, in: Proceedings of the 15th international conference on World Wide Web, ACM, 2006, pp. 633–642.

[54] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1183–1192.

[55] K. Wang, D. Zhang, Y. Li, R. Zhang, L. Lin, Cost-effective active learning for deep image classification, IEEE Transactions on Circuits and Systems for Video Technology 27 (12) (2016) 2591–2600.

[56] X. Cao, B. Qiu, X. Li, Z. Shi, G. Xu, J. Xu, Multidimensional balance-based cluster boundary detection for high-dimensional data, IEEE transactions on neural networks and learning systems 30 (6) (2018) 1867–1880.

[57] X. Cao, A divide-and-conquer approach to geometric sampling for active learning, Expert Systems with Applications (2019) 112907.

[58] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572.