ISyE 6413: Design and Analysis of Experiments

Sequential Design / Active Learning For Classification Problem

Team Members: Yishin Gan and Pouya Ahadi

4 Dec, 2022

Contents

1 Introduction **1**

2 Methodology **2**

2.1 Uncertainty Sampling For Query Selection 3

2.2 Drawback of Uncertainty Sampling 4

2.3 Space Filling Designs For Active Learning 4

3 Experiments and Results **5**

3.1 Experiments and Results For Uncertainty Sampling 5

3.2 Experiments and Results For AL with Exploration 6

4 Conclusion **7**

1. Introduction

The classification problem is one of the fundamental problems in machine learning and applied statistics areas. It lies within the *supervised learning* problem where the available data includes features and their corresponding labels. There are many classification algorithms that are already developed for this problem, and especially the boosting approach achieves much attention for the purpose of classifying. For all of these algorithms, the labels of data points are the crucial input. Thus, having a fully labeled data set is an ideal case for our purpose. However, annotating (labeling) samples requires time-consuming or expensive in many applications.

As an example, Singh et al. (2009) studies the image classification problem where image annotation can be time-consuming. As another example, in additive manufacturing, label acquisition is becoming time-consuming and labor-intensive due to sensor integration of some high-dimensional large data sets (van Houtum and Vlasea, 2021). Given the high volume of data sets in many of these applications, it turns out that annotating all instances is a very difficult or even impossible process. Thus, the training process should be done with limited labeled data. Semi-supervised algorithms use both labeled and unlabeled data to train the machine learning models. However, those algorithms do not guide decision-makers in choosing which data points to annotate. The question of "which data points should be selected for annotation?" is answered by some sequential design approaches.

Sequential experimental design can be utilized in classification problems to over with the limited resources for annotation (Wang et al., 2014). The sequential process used within the classification problem is often called *active learning (AL)*. AL is a sequential process where the user utilizes *adaptive sampling* strategies to choose the best samples for annotation. It has been shown that the uncertain samples provide more information about the underlying class distribution, and hence, they provide the best candidates for AL sampling (Settles, 2009).

Uncertainty sampling has been widely used for query selection in AL. However, under some conditions, it can be unsuccessful in training a proper classifier. Kazerouni et al. (2020) explained the conditions and conducted experiments to show that uncertainty sampling approaches tend to *exploit*. Thus, the exploitation-exploitation dilemma is an important role throughout AL, and using exploration approaches will improve the performance of the classifier.

In this project, we study and implement sequential design for the classification problem using the AL process. We utilize the exploitation-exploration trade-off and perform some experiments. As a new approach, we use the max-min distance design for the exploration and evaluate this approach.

2. Methodology

Active learning for classification problems is an iterative process where each iteration or cycle includes a sampling step and a training step. The overall procedure for AL is shown in Figure1. A typical AL algorithm starts with a base classifier. This classifier is a naive classifier trained using a small proportion of initially labeled data points. The initially labeled data can either come from historical data or be achieved by random sampling from space.

In the next step, we choose samples from the pool of unlabeled data points. The sampling criteria can be either done randomly, as it is called *random sampling*, or by using a sampling function using the information of the classifier. Based on the budget limitation, we can only sample a limited number of data points at each cycle. We denote the sampling budget as B. The sampled data points are also referred to as *query point*. Once the query points are selected, we receive the corresponding labels and retrain the classifier with new labels. The new classifier is used in the next cycle, and this procedure is repeated for a limited number of cycles based on budget.

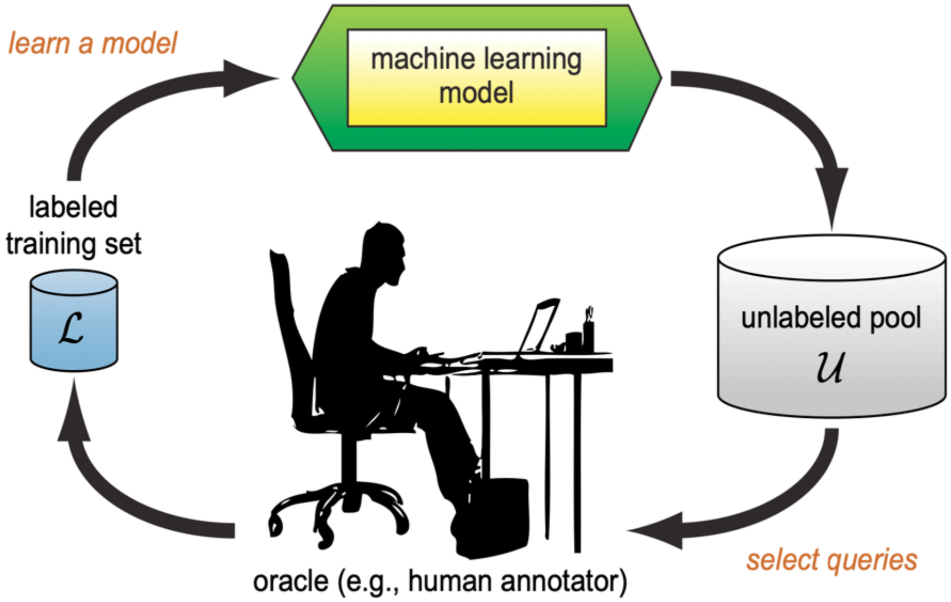


Figure 1: Procedure For AL in Classification

2.1 Uncertainty Sampling For Query Selection

Query selection is the most important step in the AL problem. As it was mentioned before, query points can be either selected randomly or by using some sampling functions. One of the common approaches for finding query points is to use *uncertainty sampling* approaches. Uncertainty sampling methods find the most informative data points by exploiting the information from the classifier Settles (2009).

It is shown that the data points closer to the decision boundary of the classifier are *approximately* most informative points. Here emphasize the word *approximately* since each classifier that we use to extract information is a naive classifier trained based on the proportion of data, and hence the information should be inaccurate but still useful for the design.

In a specific cycle of AL, using the most updated classifier, we can estimate class probabilities of a given unlabeled data point x. More specifically, $\mathbb{P}(y_i \mid x)$ is the estimated probability that data point x belongs to class y_i. Given this definition, there are three typical uncertainty sampling approaches where U is the set of unlabeled data points:

1. Least Confidence: x^*_{LC} is the query point where,

$$x^*_{LC} = \underset{x \in U}{\arg\max} \left(1 - \mathbb{P}(\hat{y} \mid x)\right), \tag{1}$$

   and $\hat{y} = \arg\max_{y_i} \mathbb{P}(y_i \mid x)$ is representing the class with highest probability for x.

2. Margin Sampling: x^*_M is the query point where,

$$x^*_M = \underset{x \in U}{\arg\min} \left(\mathbb{P}(\hat{y_1} \mid x) - \mathbb{P}(\hat{y_2} \mid x)\right), \tag{2}$$

   where $\mathbb{P}(\hat{y_1} \mid x)$ and $\mathbb{P}(\hat{y_2} \mid x)$ are first and second most probable classes.

3. Entropy Sampling: x^*_E is the query point where,

$$x^*_E = \underset{x \in U}{\arg\max} - \sum_i \mathbb{P}(y_i \mid x) log \mathbb{P}(y_i \mid x). \tag{3}$$

Based on uncertainty sampling, at each cycle, we evaluate all unlabeled data points based on one of the uncertainty measures and choose the top data points with the highest uncertainty. Among the uncertainty approaches, *entropy sampling* is more commonly used; hence, we will use entropy to find query points.

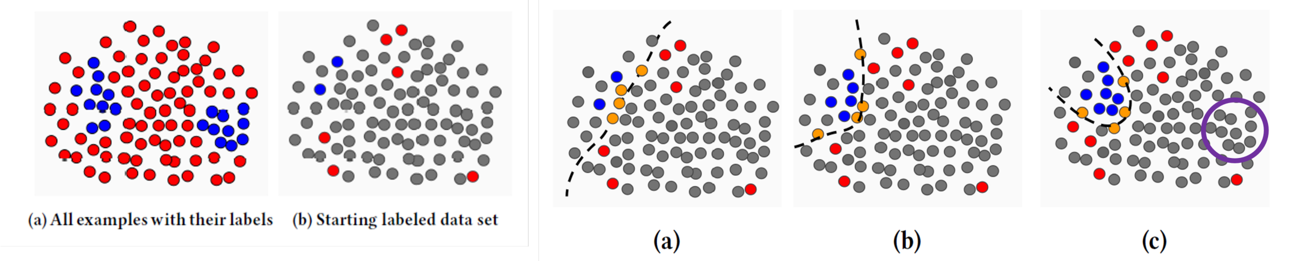
2.2 Drawback of Uncertainty Sampling

Uncertainty sampling is a powerful tool for designing AL algorithms and its widely used in many applications. However, it has been shown that it can be problematic when he data set has some properties. When dealing with complex data sets, they can be highly imbalanced which means some classes are very small compared to the majority classes. Most of the times small classes appear as small clusters in the space. Also, in many applications, due to the big data that we deal with, we have to start with a very small labeled data points due to our budget limitation which means the initially labeled data will not reflect enough information about underlying classes (Kazerouni et al., 2020).

Under these conditions, uncertainty sampling will tend to *exploit* the specific regions of data where the classifier already constructed a decision boundary. This means the AL process will keep sampling from those regions and it will be unable to detect any new decision boundary due to the lack of *exploration*.

Figure 2 represents the draw back of uncertainty sampling with an example. The data set includes two clusters of *class blue* and the initially labeled data points are not fully representing the information for underlying classes. As we can observe the AL algorithm will be able to detect the left blue cluster using uncertainty sampling but the cluster on the right side will be undetected.

[*Source: Kazerouni et al.* 2020]



(a) All examples with their labels (b) Starting labeled data set

(a) (b) (c)

Figure 2: An example representing the issue of uncertainty sampling

2.3 Space Filling Designs For Active Learning

In order to solve the issue with uncertainty sampling, we need to add *exploration* to query selection. This will be done using techniques of *space-filling designs*. In order to do so, we first divide our sampling budget into two part, $B = B_{\text{exploitation}} + B_{\text{exploration}}$. For the first share of budget, we will use entropy sampling like we mentioned before. For the second part of budget we use space0filling designs. Kazerouni et al. (2020) covered two approached for exploration: *random sampling* and *Gaussian exploration*.

Random sampling approach will choose random queries from the space. It avoids the exploitation of information however, it does not care about the coverage of the space as two random samples might be takes close to each other. Gaussian exploration approximate a exploration function using Gaussian kernels and performs better. In order to try a new approach for exploration, we implement *max-min distance design* approach for our experiment. We use following notations to formulate our model. For a specific cycle in AL we have:

1. D : set of all data points.

2. L: set of labeled data points

3. Q: set of query data points up to this iteration.

The next point to be sampled will be:

$$x^* = \underset{x \in D \setminus \{L \cup Q\}}{\arg\max} \left\{ \min_{y \in L \cup Q} \|x - y\|_2 \right\} \tag{4}$$

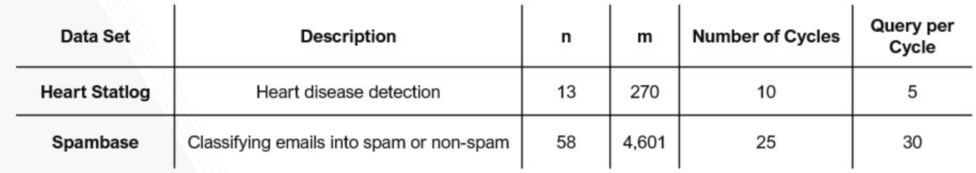
Once a sample is drawn, it will be added to set Q and we will solve Model (4) again to find the next sample. This will be repeated until we find all $B_{\text{exploration}}$ samples required for space-filling.

3. Experiments and Results

We run two types of experiments. Firstly, we conduct experiments to assess entropy sampling by comparing it to random sampling. In the other experiment, we assess our approach for exploration using a simulated data.

3.1 Experiments and Results For Uncertainty Sampling

We perform the uncertainty sampling experiment on two data sets. The data sets can be downloaded from *UCI Machine Learning Repository.* Description of data sets and the design for AL model is given in Table 3.



Data Set	Description	n	m	Number of Cycles	Query per Cycle
Heart Statlog	Heart disease detection	13	270	10	5
Spambase	Classifying emails into spam or non-spam	58	4,601	25	30

Figure 3: Description of data sets and AL designs. m represents the number of data points and n represents the number of features.

We run our experiments for 30 replications and use $F1$ score to evaluate approaches. The results are shown in Figure 4. We can observe from results that entropy sampling is more efficient in terms of exploitation and constructing a reliable classifier.

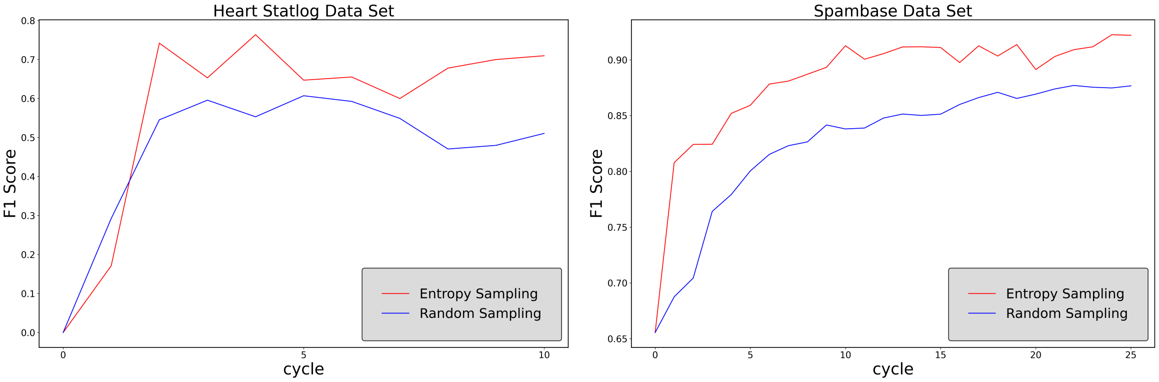


Figure 4: Comparing entropy sampling with random sampling

3.2 Experiments and Results For AL with Exploration

We simulate a two-dimensional data set with binary classes. The data set includes $20,000$ data points including 1.5% of positive samples existing within three clusters. Figure 5 represents the the data set and the results based on F1-score. We compare entropy sampling (only exploitation) with entropy sampling with max-min distance design. The sampling budget for each cycle is 100 and we use 10 samples for space-filling design.
As we can observe, the data set follows the conditions as we mentioned before, its skewed and possitive class appears with clusters. Also, data set is large and we consider 0.1% of data as initially labeled. The results shows that the entropy sampling will converge to a smaller value compared to the case where we add exploration. This will represent the effectiveness of the space-filling design and more specifically the approach that we tried, max-min distance design, in exploration data spaces and extracting unknown information or clusters from the data.

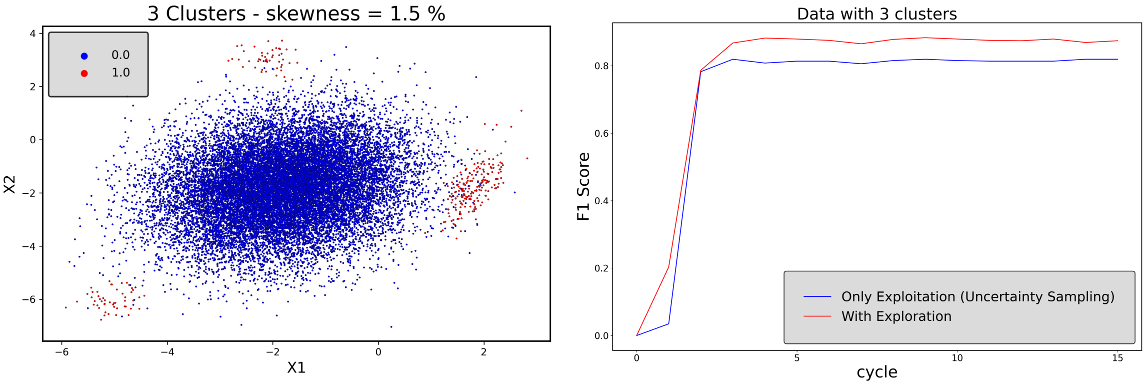


Figure 5: Effect of adding exploration on performance of AL algorithms

4. Conclusion

Due to the big data challenge in many applications, active learning approaches gain more attention. On the other hand, data annotation is becoming challenging due to resource limits. The main part of an AL algorithm sampling dilemma and how to choose representative samples to construct a reliable machine learning models using limited samples.

Designing a sampling approach might be challenging due to the complex structure of underlying class distributions. Uncertainty sampling approaches are most common approaches. Our experiments on two real data sets shows the fact that these approaches can be useful compared to a baseline method like random sampling.

On the other hand, complexity of some data sets requires use of exploration to detect all useful information. Space-filling design approaches can be utilized to cover the data space and identify information's. Our experiment shows how max-min distance design can be used as an exploration technique and lead the AL procedure to construct a reliable classifier.

References

Kazerouni, A., Zhao, Q., Xie, J., Tata, S., and Najork, M. (2020). Active learning for skewed data sets. *arXiv preprint arXiv:2005.11442*.

Settles, B. (2009). Active learning literature survey.

Singh, M., Curran, E., and Cunningham, P. (2009). Active learning for multi-label image annotation. Technical report, University College Dublin. School of Computer Science and Informatics.

van Houtum, G. J. and Vlasea, M. L. (2021). Active learning via adaptive weighted uncertainty sampling applied to additive manufacturing. *Additive Manufacturing*, 48:102411.

Wang, J., Park, E., and Chang, Y.-c. I. (2014). Active learning via sequential design and uncertainty sampling.