Active Model Selection: A Variance Minimization Approach

Mitsuru Matsuura MATSUURA326@AR.SANKEN.OSAKA-U.AC.JP

Satoshi Hara SATOHARA@AR.SANKEN.OSAKA-U.AC.JP

SANKEN, Osaka Univercity
Mihogaoka, Ibaraki, Osaka

Abstract

The cost of labeling is a significant challenge in practical machine learning. This issue arises not only during the learning phase but also at the model evaluation phase, as there is a need for a substantial amount of labeled test data in addition to the training data. In this study, we address the challenge of active model selection with the goal of minimizing labeling costs for choosing the best-performing model from a set of model candidates. Based on an appropriate test loss estimator, we propose an adaptive labeling strategy that can estimate the difference of test losses with small variance, thereby enabling the estimation of the best model using fewer labeling cost. Experimental results on real-world datasets confirm that our method efficiently selects the best model.

Keywords: Active learning, model selection

1. Introduction

The labeling cost is a crucial problem in practical machine learning. To train highly accurate models, it is desired to collect a huge amount of labeled data. The problem of labeling cost triggered several active research fields including active learning (Settles, 2009; Ren et al., 2021), transfer learning (Yang et al., 2020; Zhuang et al., 2020), semi-supervised learning (Van Engelen and Hoos, 2020), weakly-supervised learning (Sugiyama et al., 2022), and foundation models (Bommasani et al., 2021). These studies aim at training accurate models with limited amount of labeling, e.g., by actively selecting data instances to be labeled, or by fine-tuning existing models with a small amount of labeled data.

Evaluation of trained machine learning models is also an inevitable step when deploying models to real world use. If the models do not exhibit desirable accuracy for practical use, such models should not be deployed. An important point here is that the problem of labeling cost also emerges at this evaluation stage because we need a certain amount of labeled test data apart from the training data. Despite the importance of the evaluation stage, the labeling cost of evaluation is not widely explored, except for a few seminal works (Sawade et al., 2010; Katariya et al., 2012; Kumar and Raj, 2018; Kossen et al., 2021).

In this study, we tackle the problem of *active model selection* with an aim of reducing the labeling cost for model selection (Sawade et al., 2012). In many machine learning tasks, it is common to train a few models with different hyper-parameters and select the one with the highest performance. A typical approach for model selection is to evaluate models on a labeled test data collected independently from the training data. Here, the problem of labeling cost arises if we need a large amount of labeled test data for model selection. To

relieve the labeling cost, we propose an algorithm for actively selecting data instance to be labeled so that we can identify the best-performing model with a small amount of labeling.

2. Active Testing

In this section, we review the framework of active testing proposed by Kossen et al. (2021), which constitutes the basis of our study. The goal of active testing is to estimate the average test loss $R = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), y_i)$ by labeling only on the subset of $D_{\mathrm{U}} = \{x_i\}_{i=1}^{N}$ so that we can reduce the labeling cost. Kossen et al. (2021) proposed an active strategy that selects one test instance to be labeled at a time. Here, ℓ is a loss function the users want to evaluate, which takes the predictions $f(x_i)$ of model f and the true labels y_i as input and returns the loss. Below, we use the shorthand notation $\ell_i \coloneqq \ell(f(x_i), y_i)$

There are two proposals in active testing. The first proposal is using *Leveled Unbiased Risk Estimator* (LURE) (Farquhar et al., 2021) to estimate R given the actively labeled test instances. Suppose one obtained labels $y_{i_1}, y_{i_2}, \ldots, y_{i_M}$ for M sequentially selected points $x_{i_1}, x_{i_2}, \ldots, x_{i_M}$ using some query distributions $q_1, q_2, \ldots, q_M$. Then, LURE is given as

$$L(M) = \frac{1}{M}\sum_{m=1}^{M} v_m^M \ell_{i_m}, \quad v_m^M = 1 + \frac{N-M}{N-m}\left(\frac{1}{(N-m+1)q_m(i_m)} - 1\right), \tag{1}$$

where $q_m(i_m)$ represents the probability of selecting $i_m \in U_{m-1}$ at the m-th step when $i_{1:m-1} = \{i_1, i_2, \ldots, i_{m-1}\}$ have been labeled. It is known that $L(M)$ is unbiased, i.e., $\mathbb{E}_{i_{1:M}}[L(M)] = R$ for any M.

The second proposal is considering the sampling distribution q_m so that the variance of LURE to be minimized. For the unbiased estimator $L(M)$, the question is how large it can deviate from the true value R. If $L(M)$ has a small variance, it ensures $L(M)$ resides in the neighborhood of R with high probability. Thus, $L(M)$ with small variance is ideal. The ideal distribution q_m^* minimizing the variance (Farquhar et al., 2021) is given by $q_m^*(i) \propto \ell_i, \forall i \in U_{m-1}$. Apparently, this ideal distribution is unavailable for the data points that are not yet labeled. Kossen et al. (2021) therefore proposed an approximation $q_m^{\mathrm{AT}}(i) \propto \mathbb{E}_{\pi(z|x_i)}\left[\ell(f(x_i), z)\right], \forall i \in U_{m-1}$, where π is a surrogate model[1] approximating the labeling mechanism for any given input x. The surrogate model is trained, e.g., by using the same training set of f.

3. Proposed Method for Active Model Selection

We now formalize the problem of active model selection (Sawade et al., 2012). We then propose an algorithm for active model selection by extending the framework of active testing. Let $f_1, f_2, \ldots, f_K$ be K model candidates, and $D_{\mathrm{U}} = \{x_i\}_{i=1}^{N}$ be a set of unlabeled test instances. If we know the true label y_i corresponding to x_i for all the instances, we can compute the average test loss$R_k = \frac{1}{N}\sum_{i=1}^{N} \ell(f_k(x_i), y_i)$. Then, we can find the model with the minimum loss[2] as $k^* = \mathrm{argmin}_k\, R_k$. Of course, this is possible only in an ideal situation where all the labels are known. The goal of active model selection is to estimate k^* by a limited amount of labeling.

1. For the choice of the surrogate π, see Kossen et al. (2021) for the detailed discussions and experiments.
2. When there are multiple models with the same minimum loss, it is sufficient to find only one of them.

3.1 Proposed Method for $K = 2$

We first consider a simple setting where the number of model candidates $K = 2$. In this situation, the model selection problem reduces to the problem of estimating the sign of the difference $R_1 - R_2$. When $R_1 - R_2 < 0$, we can conclude that the model f_1 has a smaller average loss and hence $k^* = 1$, and $k^* = 2$ otherwise. Below, we denote $\ell_i^k \coloneqq \ell(f_k(x_i), y_i)$ and $\Delta_i \coloneqq \ell_i^1 - \ell_i^2$ for simplicity.

We first note that we can use LURE for estimating the difference $R_1 - R_2$. Indeed, for any query distribution q_m, we have $L_1(M) - L_2(M) = \frac{1}{M} \sum_{m=1}^{M} v_m^M \Delta_{i_m}$, where $L_1(M)$ and $L_2(M)$ denote LURE for the models f_1 and f_2, respectively. This is an unbiased estimator of $R_1 - R_2$, i.e., $\mathbb{E}_{i_{1:M}}[L_1(M) - L_2(M)] = R_1 - R_2$. The remaining task is to design an appropriate query distribution q_m.

We derive the ideal query distribution q_m minimizing the variance of $L_1(M) - L_2(M)$. If the estimator $L_1(M) - L_2(M)$ has a small variance, we can expect $L_1(M) - L_2(M)$ to belong in close neighborhood of $R_1 - R_2$, which allows us to estimate the sign of $R_1 - R_2$ confidently. Suppose that $i_{1:m-1}$ is given. Then, the ideal query distribution minimizing the conditional variance $\mathbb{V}_{i_m \sim q_m}[L_1(m) - L_2(m) \mid i_{1:m-1}]$ is given by[3]

$$q_m^*(i) \propto |\Delta_i|, \quad \forall i \in U_{m-1}. \tag{2}$$

The ideal distribution q_m in (2) is available only when the true label is known so that one can compute the true $|\Delta_i|$. We follow the idea of Kossen et al. (2021) and use a surrogate model π to estimate $|\Delta_i|$. The resulting query distribution is then given by

$$q_m^2(i) \propto |\hat{\Delta}_i|, \quad |\hat{\Delta}_i| = \mathbb{E}_{\pi(z|x_i)}\left[|\ell(f_1(x_i), z) - \ell(f_2(x_i), z)|\right]. \tag{3}$$

3.2 Proposed Method for General K

We extend Algorithm to the general case when the number of model candidates $K \geq 2$. Similar to the case of $K = 2$, we design the query distribution q_m minimizing the variance. The difference from $K = 2$ is that there are $\frac{K(K-1)}{2}$ combinations of the models to be compared with. The question here is that which variance we aim to minimize.

In our proposed method, we consider minimizing the sum of all the pairwise variances, which is given by $V \coloneqq \sum_{k<k'} \mathbb{V}_{i_m \sim q_m}[L_k(m) - L_{k'}(m) \mid i_{1:m-1}]$. Suppose that $i_{1:m-1}$ is given. Then, the query distribution minimizing the sum of

Algorithm 1 Proposed Model Selection

Input: Models $\{f_k\}_{k=1}^K$, Surrogate π, Unlabeled data D_{U}

Output: Estimated optimal model index $\hat{k}$

$U_0 \leftarrow \{1, 2, \ldots, N\}$, $Q_0 \leftarrow \emptyset$
for $m = 1$ to M **do**
  Calculate $q_m^K(i)$ in (5) for all $i \in U_{m-1}$.
  Sample $i_m \sim q_m^K(i)$ and observe y_{i_m}.
  $U_m \leftarrow U_{m-1} \setminus \{i_m\}$, $Q_m \leftarrow Q_{m-1} \cup \{(\ell_{i_m}^1, \ell_{i_m}^2, \ldots, \ell_{i_m}^K, q_m^K(i_m))\}$.
end for
Calculate $L_k(M)$ by using Q_M for $k = 1, 2, \ldots, K$.
return $\mathrm{argmin}_{k \in \{1,2,\ldots,K\}} L_k(M)$.

3. See Appendix A for the proof.

all the pairwise variances V is[4]

$$q_m^*(i) \propto \Delta_i^V \coloneqq \sum_{k<k'} \left|\Delta_i^{k,k'}\right|, \ \forall i \in U_{m-1}, \quad \text{where } \Delta_i^{k,k'} \coloneqq \ell_i^k - \ell_i^{k'}. \tag{4}$$

Similar to the case of $K = 2$, we approximate $|\Delta_i^V|$ using the surrogate π. The resulting sampling distribution is then given by

$$q_m^K(i) \propto |\hat{\Delta}_i^V|, \quad |\hat{\Delta}_i^V| = \sum_{k<k'} \mathbb{E}_{\pi(z|x_i)} \left[|\ell(f_k(x_i), z) - \ell(f_{k'}(x_i), z)|\right]. \tag{5}$$

The pseudo-code of the proposed algorithm for general K is shown in Algorithm 1.

4. Related Work

The objective of active model selection is to estimate the best-performing model from the candidates by a small number of labeling. The study most close to ours would be Sawade et al. (2012). They proposed to estimate the average test loss using the importance weighting

$$L_k^{\mathrm{W}}(M) = \frac{1}{W_M} \sum_{m=1}^{M} \frac{1}{q_m(x_{i_m})} \ell_{i_m}^k, \quad W_M = \sum_{m=1}^{M} \frac{1}{q_m(x_{i_m})}. \tag{6}$$

They showed that the ideal query distribution minimizing the asymptotic variance of $L_1^{\mathrm{W}}(M) - L_2^{\mathrm{W}}(M)$ is $q_m^*(i) \propto \sqrt{\mathbb{E}_{p(z|x_i)}\left[\left(\Delta_\ell(x_i, z) - \Delta_R\right)^2\right]}$, where $\Delta_\ell(x, z) \coloneqq \ell(f_1(x), y) - \ell(f_2(x), z)$ and $\Delta_R \coloneqq R_1 - R_2$.

There is one essential difference with the current study and Sawade et al. (2012). Our query distribution (2) minimizes the conditional variance of the estimated test loss in each step, while Sawade et al. (2012) minimizes the asymptotic variance only. As we show in the experiments in Section 5, this difference is significant. The proposed method incurs small variance even for a small number of labeling M, while the method of Sawade et al. (2012) tends to incur large variance for small M.

5. Experiment

In this section, we demonstrate the effectiveness of the proposed method in two ways. First, we show that the proposed method can estimate the difference of test losses with small variance, so that we can confirm our theoretical claim. Second, we demonstrate that the proposed method can estimate the best-performing model with a small number of labeling compared to the existing baselines.

5.1 Setups

Datasets & Loss We used the four datasets obtained from LIBSVM Data repository[5] shown in Table 1. In the experiments, we randomly subsampled 5,000 instances as the training set and subsampled another 500 instances as the test set. Because these are the classification datasets, we used the zero–one loss as the loss function ℓ.

4. See Appendix A for the proof.
5. `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html`

As the multiple model candidates $f_1, f_2, \ldots, f_K$, we used four-layer perceptron (MLP) with different number of hidden neurons. We trained each model f_k on the training set using `MLPClassifier` of scikit-learn with `hidden_layer_sizes` = $(b, b/2)$ for several different values of b and `max_iter` = 1000. We used the default values for the other options.

dataset	classes	features	instances
covtype	7	54	581,012
letter	26	16	15,000
mnist	10	780	60,000
sensorless	48	11	58,509

Table 1: Datasets used for the experiments.

As the surrogate π, we used the ensemble $\frac{1}{K}\sum_{k=1}^{K} f_k$ proposed by Sawade et al. (2012).

Baselines We compared the proposed method (PROPOSED) with the two baselines. The first baseline is UNIFORM that selects the next labeling instance uniformly at random. This is the most naive baseline. The second baseline is SAWADE using the loss estimator and the query distribution proposed by Sawade et al. (2012).

5.2 Experiment 1: Comparison of Variance

We first validate the variance of the difference of estimated test losses. Here, we used MLP of $b = 100$ as the first model f_1 and MLP of $b = 1000$ as the second model f_2. Table 2 shows the test losses of the models.

	covtype	letter	mnist	sensorless
R_1	0.238	**0.092**	0.082	**0.028**
R_2	**0.228**	0.124	**0.074**	0.030

Table 2: [Experiment 1] The average test losses R_1 and R_2 for MLPs f_1 with the number of neurons $b = 100$ and f_2 with $b = 1000$, respectively. The bold loss denote R_{k^*} corresponding to the best model with the minimum test loss for each dataset.

Figure 1 shows the results of active model selection. We run each method of UNIFORM, SAWADE, and PROPOSED for 1000 times with different random seeds. Figure 1 shows the average and the standard deviation of the estimated test loss difference $\Delta_R = R_{k^*} - R_{k_2^*}$ over the number of labeling M, where R_{k^*} and $R_{k_2^*}$ denote the minimum and the second minimum test loss.

It is evident from the figure that PROPOSED attained the smallest standard deviation (std.) for any M compared to the baselines. The figures show that PROPOSED converged to the true Δ_R around $M = 100$. Moreover, we can see that its upper line (average + std.) gets smaller than zero (black dashed line) for very small $M = 20 \sim 50$. That is, PROPOSED could estimate the sign of the difference and thereby identifying the best model after labeling $M = 20 \sim 50$ instances most of the times.

	covtype	letter	mnist	sensorless
R_1	0.262	0.160	0.092	0.034
R_2	0.298	0.110	0.078	0.036
R_3	<u>0.238</u>	<u>0.092</u>	0.082	**0.028**
R_4	0.248	**0.068**	<u>0.074</u>	**0.028**
R_5	0.246	<u>0.092</u>	**0.068**	**0.028**
R_6	**0.228**	0.124	<u>0.074</u>	<u>0.030</u>

Table 3: [Experiment 2] The average test losses $R_1, R_2, \ldots, R_6$ for MLPs $f_1, f_2, \ldots, f_6$ with the number of neurons $b = 30, 50, 100, 300, 500, 1000$, respectively. The bold and underlined losses denote R_{k^*} and $R_{k_2^*}$, respectively, corresponding to the minimum and second minimum test losses for each dataset.

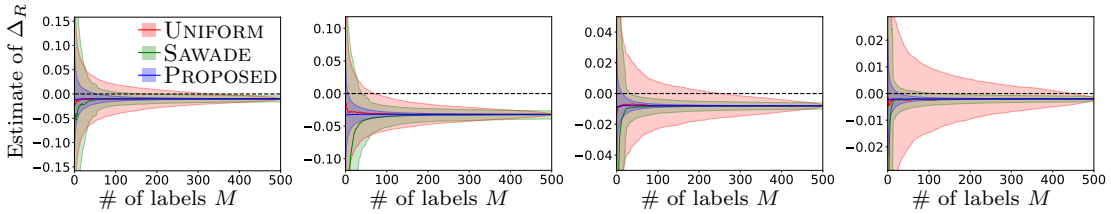


Figure 1: [Experiment 1] The estimated test loss differences over the number of labels M. The average estimates are drawn in solid line while their standard deviations (average $\pm$ std.) are denoted by the shaded regions.

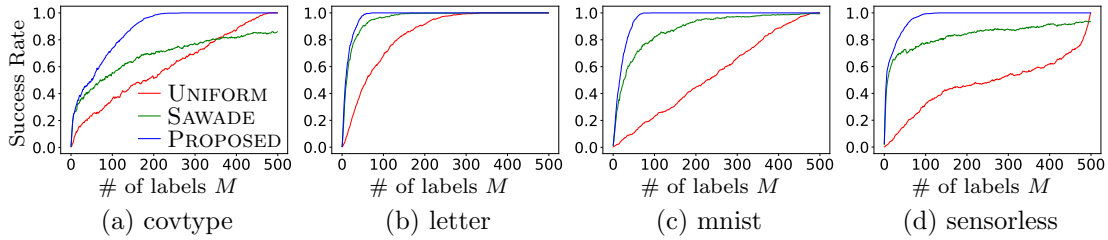


Figure 2: [Experiment 2] The success rates of identifying the best model over the number of labels M.

5.3 Experiment 2: Estimation of the Best-Performing Model

We now demonstrate that the proposed method can find the best-performing model effectively. Here, we used MLP of $b = 30, 50, 100, 300, 500, 1000$ as the model candidates $f_1, f_2, f_3, f_4, f_5, f_6$, shown in Table 3. We run each method of UNIFORM, SAWADE, and PROPOSED for 1000 times with different random seeds.

Figure 2 shows the success rate of each method after labeling M instances, i.e., the ratio of the cases when the method can successfully identify the best model over the 1000 trials. In the figures, we can see that PROPOSED dominates the other baselines; it could approach to the perfect success rate with the smaller labeling cost. In particular, for the letter, mnist, and sensorless datasets, it attained the perfect success rate with $M \approx 100$, far smaller cost than the baselines. The results also confirm that UNIFORM tends to be the worst most of the cases as expected. SAWADE was far better than UNIFORM showing its effectiveness for active model selection.

6. Conclusion

In this study, we tackled the problem of active model selection with an aim of reducing the labeling cost for selecting the best-performing model. We formulated the problem as the estimation of the sign of the difference of test losses. To solve the problem, we derived the ideal query distribution that minimizes the variance of the estimated test loss difference, so that one can estimate the best model with confidence. Experiments on real-world datasets confirmed that our method can estimate the difference of test losses with small variance, leading to an effective model selection with a small number of labeling.

Acknowledgments

We would like to thank Dr. Junya Honda for useful discussions and valuable feedbacks. SH was supported by JSPS KAKENHI Grant Number 20K19860, 23H03456, and JST, PRESTO Grant Number JPMJPR20C8.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Namit Katariya, Arun Iyer, and Sunita Sarawagi. Active evaluation of classifiers on large datasets. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 329–338, 2012.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5753–5763, 2021.

Anurag Kumar and Bhiksha Raj. Classifier risk estimation under limited labeling resources. In *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–15, 2018.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning*, pages 951–958, 2010.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. In *Advances in Neural Information Processing Systems*, 2012.

Burr Settles. Active learning literature survey. 2009.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach*. MIT Press, 2022.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. *Transfer Learning*. Cambridge University Press, 2020.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

Appendix A. Proof.

In this appendix, we prove the following theorems from Sections 3.1 and 3.2.

Theorem 1 (Query Distribution with Minimum Variance) *Suppose that $i_{1:m-1}$ is given. Then, the query distribution minimizing the conditional variance $\mathbb{V}_{i_m \sim q_m}\left[L_1(m) - L_2(m) \mid i_{1:m-1}\right]$ is given by*

$$q_m^*(i) \propto |\Delta_i|, \quad \forall i \in U_{m-1}. \tag{2}$$

Proof We recall that, by the definition of LURE and v_m^m,

$$L_1(m) - L_2(m) = \frac{1}{m}\sum_{j=1}^{m-1} v_j^m \Delta_{i_j} + \frac{1}{m}\frac{1}{N-m+1}\frac{\Delta_{i_m}}{q_m(i_m)}.$$

Because $i_{1:m-1}$ is given and fixed, only the last term is the random variable whose expectation is $\mathbb{E}_{i_m \sim q_m}\left[\frac{\Delta_{i_m}}{q_m(i_m)}\right] = \sum_{j \in U_{m-1}} \Delta_j$. The conditional variance is then expressed as

$$\begin{aligned}
&\mathbb{V}_{i_m \sim q_m}\left[L_1(m) - L_2(m) \mid i_{1:m-1}\right] \\
&= \frac{1}{m^2}\frac{1}{(N-m+1)^2}\mathbb{E}_{i_m \sim q_m}\left[\left\{\frac{\Delta_{i_m}}{q_m(i_m)} - \sum_{j \in U_{m-1}} \Delta_j\right\}^2\right] \\
&= \frac{1}{m^2}\frac{1}{(N-m+1)^2}\left[\sum_{j \in U_{m-1}} \frac{\Delta_j^2}{q_m(j)} - \left\{\sum_{j \in U_{m-1}} \Delta_j\right\}^2\right].
\end{aligned}$$

Thus, the minimization of the conditional variance is reduced to the following optimization problem.

$$\min_{q_m} \sum_{j \in U_{m-1}} \frac{\Delta_j^2}{q_m(j)}, \quad \text{s.t.} \quad q_m(j) \geq 0, \sum_{j \in U_{m-1}} q_m(j) = 1.$$

We can solve this problem by using the method of Lagrange multipliers, and the claim follows. ■

Theorem 2 (Query Distribution with Minimum V) *Suppose that $i_{1:m-1}$ is given. Then, the query distribution minimizing the sum of all the pairwise variances V is*

$$q_m^*(i) \propto \Delta_i^V \coloneqq \sum_{k<k'} \left|\Delta_i^{k,k'}\right|, \quad \forall i \in U_{m-1}, \quad where \Delta_i^{k,k'} \coloneqq \ell_i^k - \ell_i^{k'}. \tag{4}$$

Proof By the similar arguments as in the proof of Theorem 1, the minimization of V reduces to the following optimization problem.

$$\min_{q_m} \sum_{j \in U_{m-1}} \frac{\left(\Delta_j^V\right)^2}{q_m(j)}, \quad \text{s.t.} \quad q_m(j) \geq 0, \sum_{j \in U_{m-1}} q_m(j) = 1.$$

We can solve this problem by using the method of Lagrange multipliers, and the claim follows. ■

Appendix B. Experiment 1: Comparison of Variance

We show the additional results on Experiment 1 in Section 5.2.

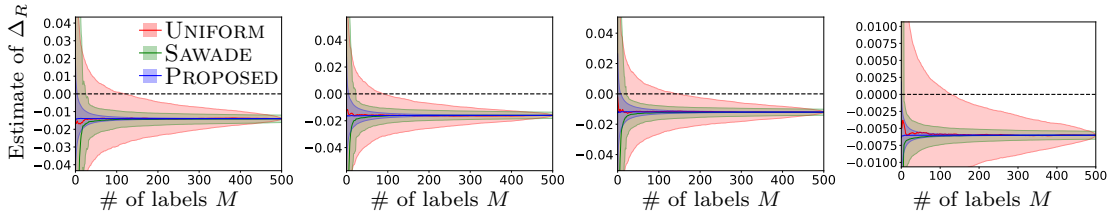
B.1 RandomForest and Ensemble Surrogate

Here, we show the results on RandomForest. As the model candidates f_1, f_2, we used RandomForest of the depth 14 and 20, respectively. We trained each model f_k using the training set using `RandomForestClassifier` of scikit-learn with `max_depth` $= b$ with $b = 14$ and $b = 20$, shown in Table 4. We used the default values for the other options. As the surrogate π, we used the ensemble $\frac{1}{2}\sum_{k=1}^{2} f_k$ proposed by Sawade et al. (2012). We run each method of Uniform, Sawade, and Proposed for 1000 times with different random seeds.

	covtype	letter	mnist	sensorless
R_1	0.226	0.104	0.082	0.012
R_2	**0.212**	**0.088**	**0.070**	**0.006**
Δ_R	- 0.014	- 0.016	- 0.012	- 0.006

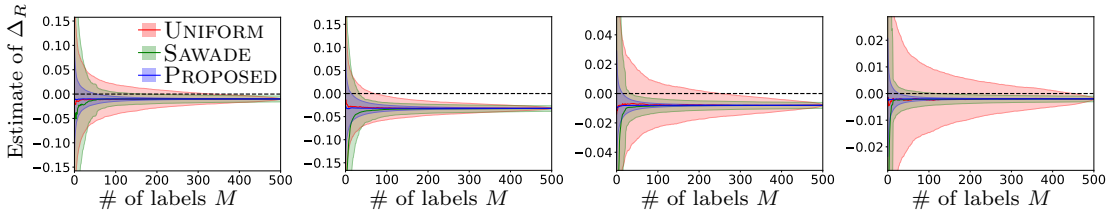
Table 4: [Experiment 1: RandomForest] The average test losses R_1 and R_2 for RandomForestss f_1 with the maximum depth $b = 14$ and f_2 with $b = 20$, respectively. The bold loss denote R_{k^*} corresponding to the best model with the minimum test loss for each dataset. $\Delta_R = R_{k^*} - R_{k_2^*}$ is the difference of test losses where $R_{k_2^*}$ denotes the second minimum test loss.

Figure 3 shows the average and the standard deviation of the estimated test loss difference $\Delta_R = R_{k^*} - R_{k_2^*}$ over the number of labeling M, where R_{k^*} and $R_{k_2^*}$ denote the minimum and the second minimum test loss. The result is similar to the case of MLP in Section 5.2 where Proposed attained the smallest variance for all M.



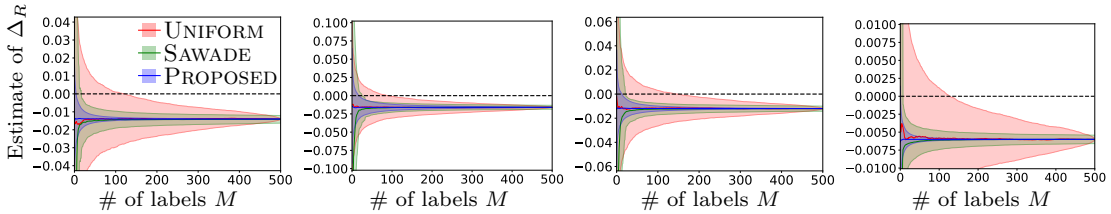
(a) covtype (b) letter (c) mnist (d) sensorless

Figure 3: [Experiment 1: RandomForest + Ensemble Surrogate] The estimated test loss differences over the number of labels M. The average estimates are drawn in solid line while their standard deviations (average ± std.) are denoted by the shaded regions.



(a) covtype (b) letter (c) mnist (d) sensorless

Figure 4: [Experiment 1: MLP + Logistic Regression Surrogate] The estimated test loss differences over the number of labels M. The average estimates are drawn in solid line while their standard deviations (average $\pm$ std.) are denoted by the shaded regions.



(a) covtype (b) letter (c) mnist (d) sensorless

Figure 5: [Experiment 1: RandomForest + Logistic Regression Surrogate] The estimated test loss differences over the number of labels M. The average estimates are drawn in solid line while their standard deviations (average $\pm$ std.) are denoted by the shaded regions.

B.2 Logistic Regression Surrogate

Here, we show the results for the case when we used the linear logistic regression as the surrogate π. We trained linear logistic regression on the training set using `LogisticRegressionCV` of scikit-learn with default options. In the experiment, we use the same model candidates, MLPs (Section 5.2, Table 2) and RandomForests (Appendix B.1, Table 4).

Figures 4 and 5 show the variance for MLP and RandomForest, respectively. In the figures, Proposed performed the best similar to the case when the ensemble surrogate is used (Section 5.2, Appendix B.1). These results show that the choice of the surrogate π did not have significant impacts to the results.

Appendix C. Experiment 2: Estimation of the Best-Performing Model

We show the additional results on Experiment 2 in Section 5.3.

C.1 RandomForest and Ensemble Surrogate

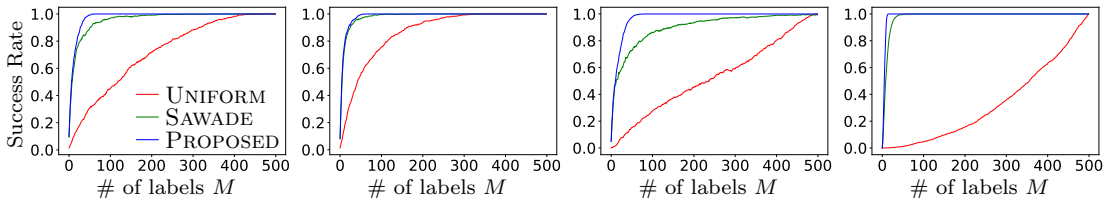
Here, we show the results on RandomForest. As the multiple model candidates $f_1, f_2, \ldots, f_K$, we used RandomForest of different depths. We trained each model f_k using the training set using `RandomForestClassifier` of scikit-learn with `max_depth` $= b$ with a prescribed value b. We used the default values for the other options. As the surrogate π, we used the ensemble $\frac{1}{K}\sum_{k=1}^{K} f_k$ proposed by Sawade et al. (2012).

	covtype	letter	mnist	sensorless
R_1	0.248	0.174	0.074	0.018
R_2	0.232	0.146	0.076	0.010
R_3	0.226	0.104	0.082	0.012
R_4	0.232	<u>0.102</u>	**0.064**	<u>0.008</u>
R_5	<u>0.222</u>	**0.088**	<u>0.068</u>	<u>0.008</u>
R_6	**0.212**	**0.088**	0.070	**0.006**
Δ_R	- 0.010	- 0.014	- 0.004	- 0.002

Table 5: [Experiment 2: RandomForest] The average test losses $R_1, R_2, \ldots, R_6$ for RandomForests $f_1, f_2, \ldots, f_6$ with the maximum depth $b = 10, 12, 14, 16, 18, 20$, respectively. The bold and underlined losses denote R_{k^*} and $R_{k_2^*}$, respectively, corresponding to the minimum and second minimum test losses for each dataset. $\Delta_R = R_{k^*} - R_{k_2^*}$ is the difference of test losses.

We used RandomfForest with the maximum depth $b = 10, 12, 14, 16, 18, 20$ as the model candidates $f_1, f_2, f_3, f_4, f_5, f_6$, shown in Table 5. We run each method of UNIFORM, SAWADE, and PROPOSED for 1000 times with different random seeds.

Figure 6 shows the success rate for identifying the best model. Similar to the results in Section 5.3, PROPOSED performed the best; it requires the small number of labeling for identifying the best model compared to the baselines. In this experiment, SAWADE was comparable with PROPOSED, although PROPOSED outperformed slightly.



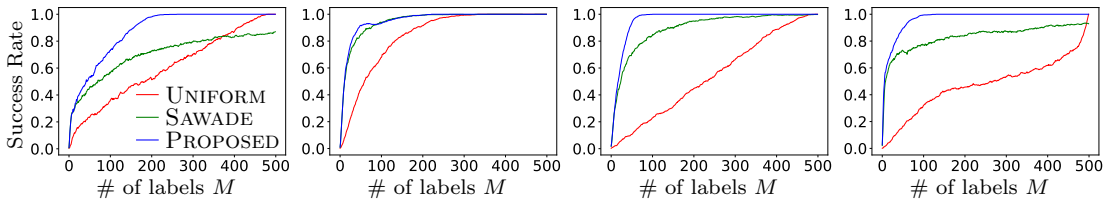
(a) covtype (b) letter (c) mnist (d) sensorless

Figure 6: [Experiment 2: RandomForest + Ensemble Surrogate] The success rates of identifying the best model over the number of labels M.

C.2 Logistic Regression Surrogate

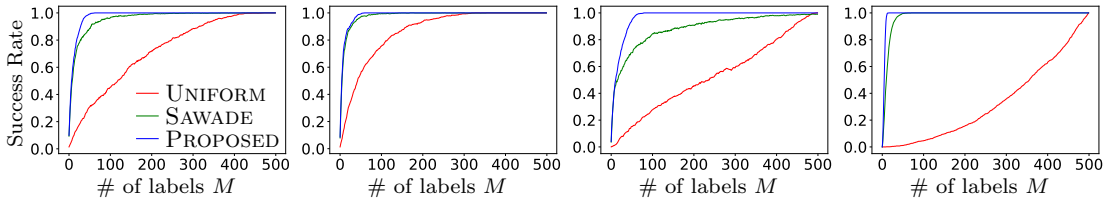
Here, we show the results for the case when used the linear logistic regression as the surrogate π. We trained linear logistic regression on the training set using `LogisticRegressionCV` of scikit-learn with default options. In the experiment, we use the same model candidates, MLPs (Section 5.3, Table 3) and RandomForests (Appendix C.1, Table 5).

Figures 7 and 8 shows the success rates for MLP and RandomForest, respectively. In the figures, PROPOSED performed the best similar to the case when the ensemble surrogate is used (Section 5.3, Appendix C.1). These results show that the choice of the surrogate π did not have significant impacts to the results.



(a) covtype (b) letter (c) mnist (d) sensorless

Figure 7: [Experiment 2: MLP + Logistic Regression Surrogate] The success rates of identifying the best model over the number of labels M.



(a) covtype (b) letter (c) mnist (d) sensorless

Figure 8: [Experiment 2: RandomForest + Logistic Regression Surrogate] The success rates of identifying the best model over the number of labels M.