Combining Self-labeling with Selective Sampling

Jędrzej Kozal, **Michał Woźniak**

Wrocław University of Science and Technology, Poland

December 1, 2023

IncrLearn - ICDM 2023

Outline

1. Motivations
2. Self Labeling Selective Sampling (SLS2)
3. Experiments
4. Conclusion

Outline

1. Motivations
2. Self Labeling Selective Sampling (SLS2)
3. Experiments
4. Conclusion

Data annotation

Unlabeled data is cheap.

Labels are expensive because:

- they may require specialized expertise (e.g., annotating one hour of recordings requires ca. 400 work hours)
- they may require specialized equipment

Data annotation

Unlabeled data is cheap.
Labels are expensive because:

- they may require specialized expertise (e.g., annotating one hour of recordings requires ca. 400 work hours)
- they may require specialized equipment

Labeling quality

Data labeling/annotation often requires dedicated techniques to eliminate or minimize human error's impact.

- In the case of medicine, misdiagnoses are estimated to account for between 3 and more than 13% of all medical decisions, [1]
- IBM estimates that the U.S. economy loses about $3 trillion a year due to poor data quality [2].

The link between the quality of learning data and models is therefore obvious.

---

[1] Linda T. Kohn, To Err is Human: Building a Safer Health System, 2000, https://pubmed.ncbi.nlm.nih.gov/25077248/

[2] Manu Bansal, Flying Blind: How Bad Data Undermines Business, https://www.forbes.com/sites/forbestechcouncil/2021/10/14/flying-blind-how-bad-data-undermines-business/?sh=63c5c7e929e8

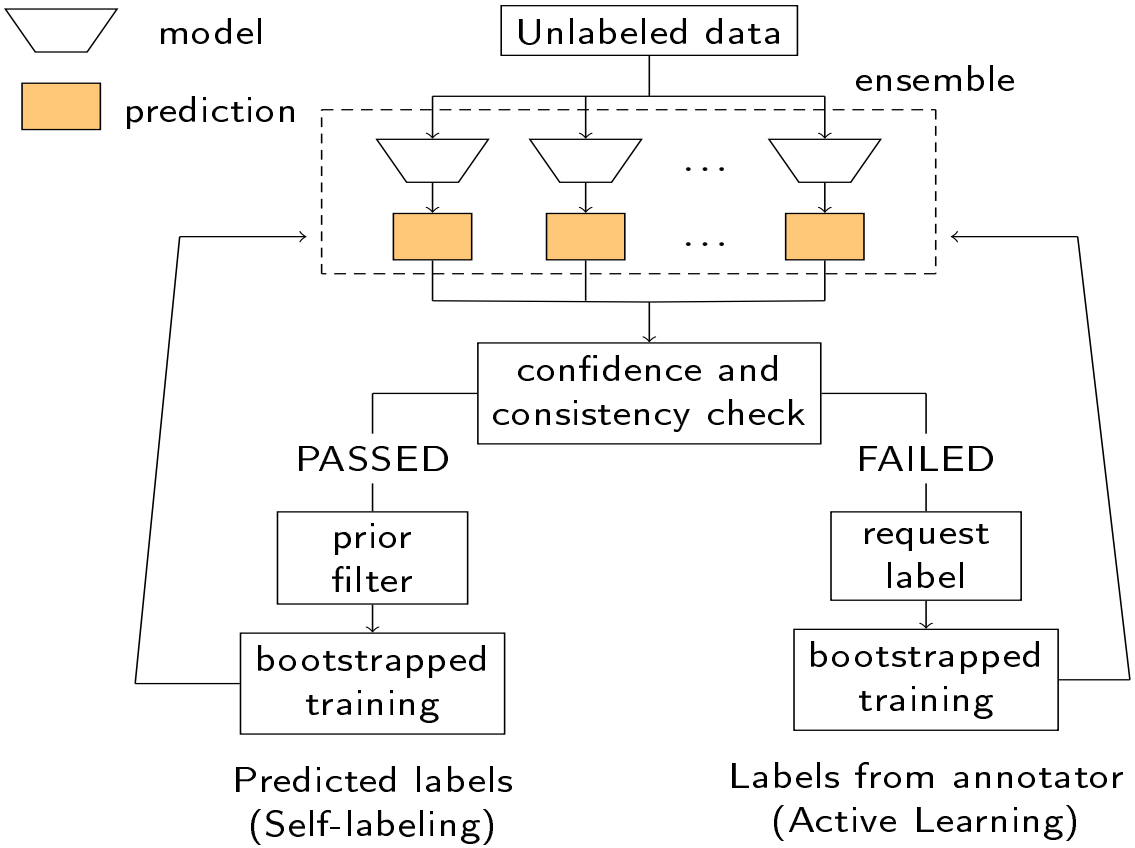
Outline

1. Motivations
2. Self Labeling Selective Sampling (SLS2)
3. Experiments
4. Conclusion

Idea

SLS2 combines self-labeling with active learning in a Stream-Based Selective Sampling scenario to minimize the cost of an expert's annotations while taking advantage of the opportunity to learn the model without the expert's participation.

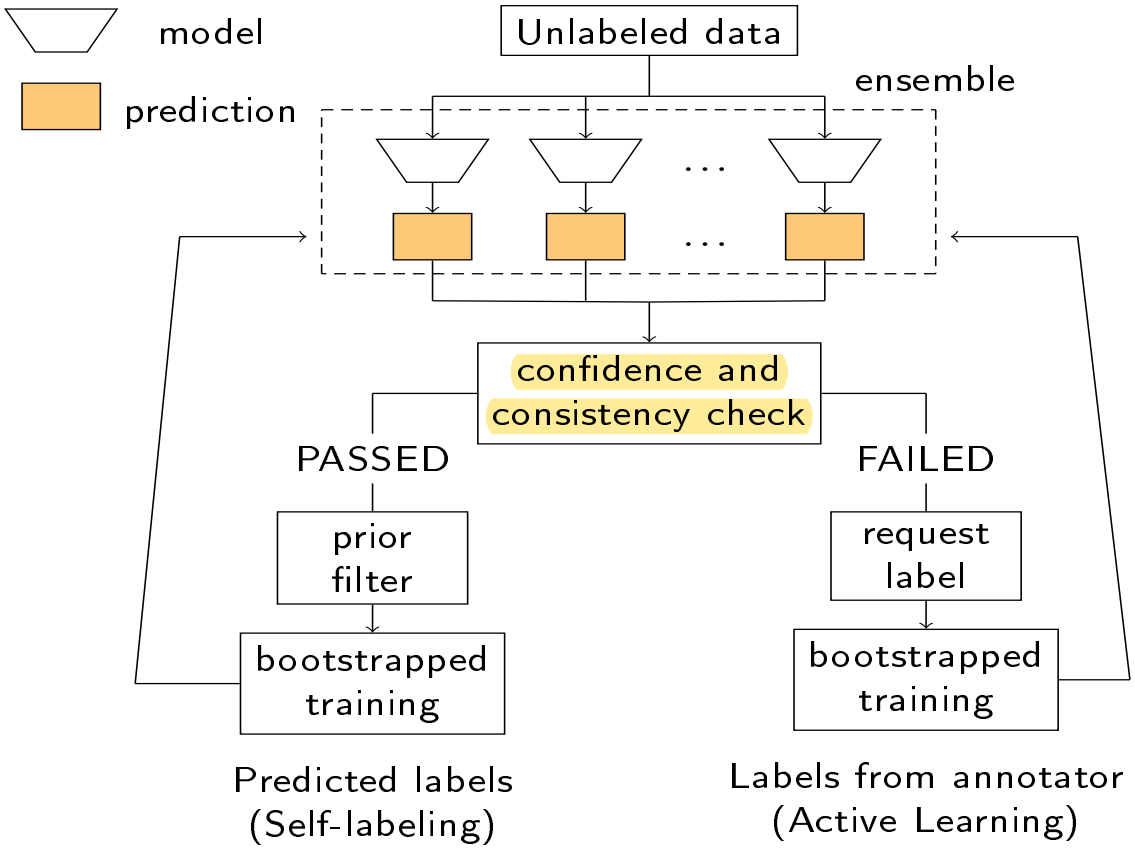
Proposed method



Selective sampling

We assume the same cost of obtaining label from an oracle for each sample. so we define budget B as the number of samples that could be labeled.

Proposed method-Informativeness computation



Informativeness computation

- For the unlabeled sample each model l computes supports $f(\boldsymbol{x}, \theta_l)$.
- Then, we check if at least half of the base classifiers provided supports that exceed a predefined threshold τ.
- If more than half of the models return confident and consistent prediction $\hat{y}$) , we add $(\boldsymbol{x}, \hat{y})$ to $\mathscr{L}$.
- Otherwise, we query an oracle with $\boldsymbol{x}$.

Informativeness computation

- For the unlabeled sample each model l computes supports $f(\boldsymbol{x}, \theta_l)$.
- Then, we check if at least half of the base classifiers provided supports that exceed a predefined threshold τ.
- If more than half of the models return confident and consistent prediction $\hat{y}$) , we add $(\boldsymbol{x}, \hat{y})$ to $\mathscr{L}$.
- Otherwise, we query an oracle with $\boldsymbol{x}$.

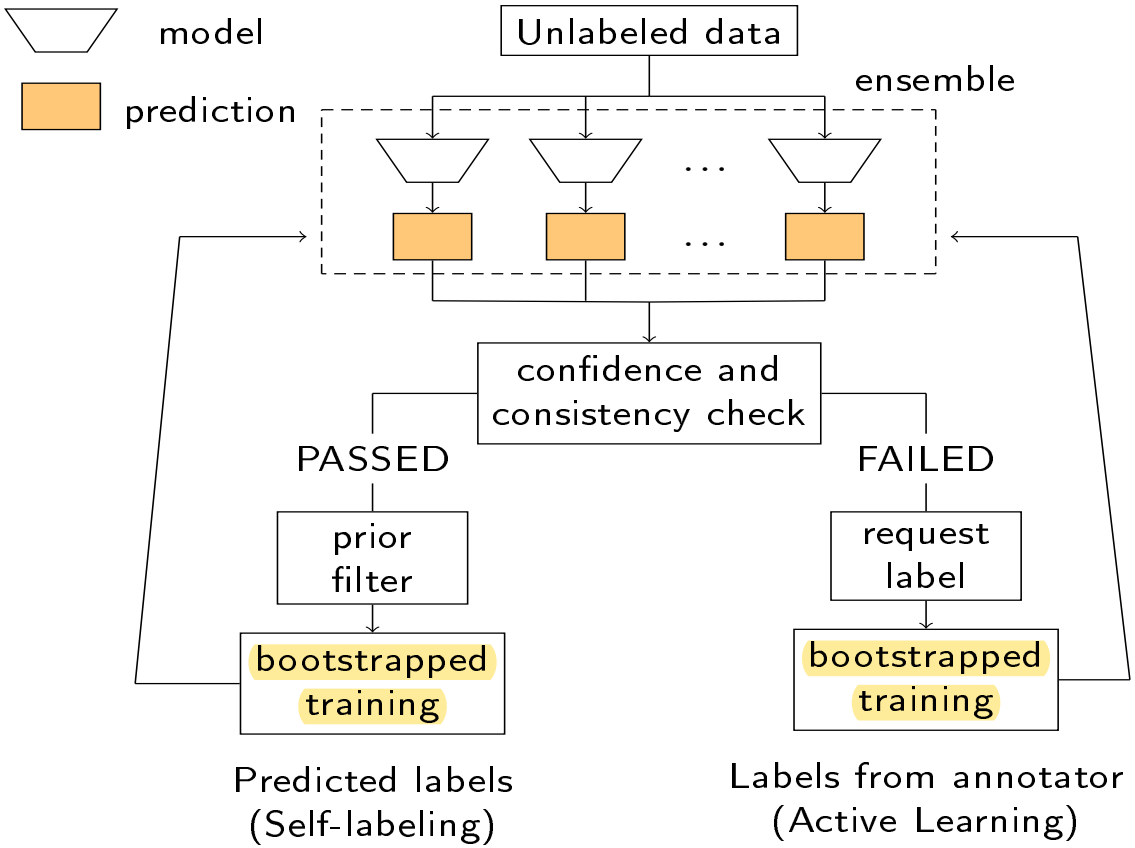
Informativeness computation

- For the unlabeled sample each model l computes supports $f(\boldsymbol{x}, \theta_l)$.
- Then, we check if at least half of the base classifiers provided supports that exceed a predefined threshold τ.
- If more than half of the models return confident and consistent prediction $\hat{y}$) , we add $(\boldsymbol{x}, \hat{y})$ to $\mathscr{L}$.
- Otherwise, we query an oracle with $\boldsymbol{x}$.

Informativeness computation

- For the unlabeled sample each model l computes supports $f(\boldsymbol{x}, \theta_l)$.
- Then, we check if at least half of the base classifiers provided supports that exceed a predefined threshold τ.
- If more than half of the models return confident and consistent prediction $\hat{y}$) , we add $(\boldsymbol{x}, \hat{y})$ to $\mathscr{L}$.
- Otherwise, we query an oracle with $\boldsymbol{x}$.

Proposed method-Bootstraped training



Bootstraped training

- We train initial models with bootstrapping of labeled part of data $\mathscr{L}$ using Online Bagging concept ($\lambda = 1$).
- In the case of training with ground truth label from oracle, we use $\lambda = 1$.
- When updating the dataset with a self-labeling λ is given by

$$\lambda = \frac{\max_{l,c} f_c(\boldsymbol{x}, \theta_l)}{\tau} - \mathbb{1}_{B=0}$$

where τ is the same threshold used earlier for selecting confident predictions (if $B > 0$, then $\lambda > 1$).

Bootstraped training

- We train initial models with bootstrapping of labeled part of data $\mathscr{L}$ using Online Bagging concept ($\lambda = 1$).
- In the case of training with ground truth label from oracle, we use $\lambda = 1$.
- When updating the dataset with a self-labeling λ is given by

$$\lambda = \frac{\max_{l,c} f_c(\boldsymbol{x}, \theta_l)}{\tau} - \mathbb{1}_{B=0}$$

where τ is the same threshold used earlier for selecting confident predictions (if $B > 0$, then $\lambda > 1$).

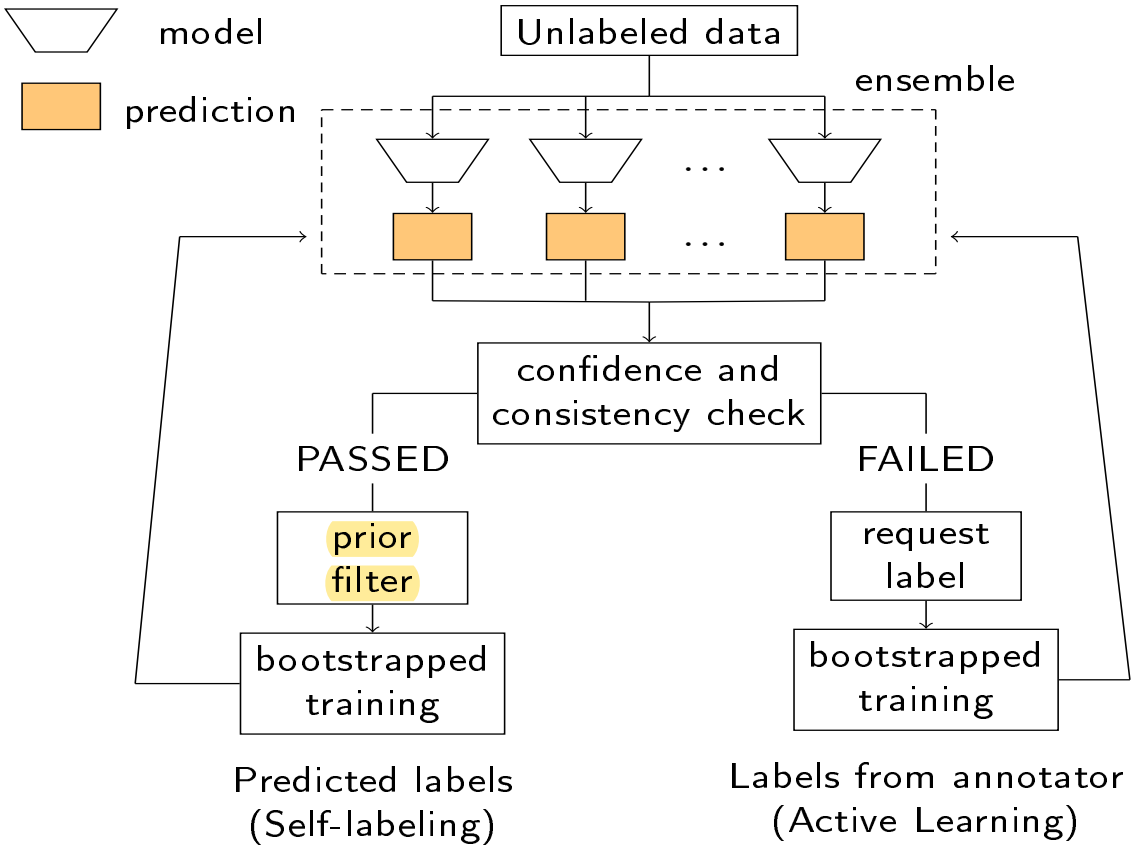
Bootstraped training

- We train initial models with bootstrapping of labeled part of data $\mathscr{L}$ using Online Bagging concept ($\lambda = 1$).
- In the case of training with ground truth label from oracle, we use $\lambda = 1$.
- When updating the dataset with a self-labeling λ is given by

$$\lambda = \frac{\max_{l,c} f_c(\boldsymbol{x}, \theta_l)}{\tau} - \mathbb{1}_{B=0}$$

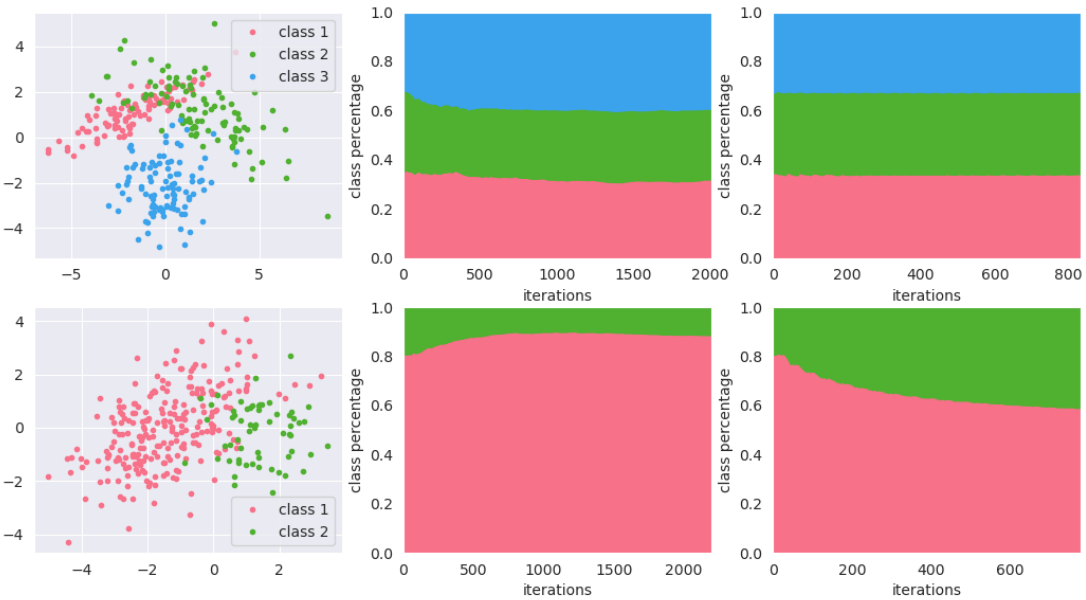
where τ is the same threshold used earlier for selecting confident predictions (if $B > 0$, then $\lambda > 1$).

Proposed method-Prior filtering



Prior filtering

This work shows that initial bias in the data distribution can be strengthened by self-labeling.
We propose simple prior filtering, which does not allow self-labeling if adding new objects would bias a class distribution.



Outline

1. Motivations
2. Self Labeling Selective Sampling (SLS2)
3. Experiments
4. Conclusion

Research questions

RQ1:Is there a benefit of combining active learning strategies with self-labeling?

RQ2: What is the performance of the proposed method for datasets with a high number of learning examples?

RQ3: How does the initial training set size (the seed size) impact the performance?

RQ4: How does the accuracy of the model trained with seed impact the learning process of the proposed algorithm?

RQ5: Does the proposed algorithm allow for the better utilization of the computational budget?

Research questions

RQ1:Is there a benefit of combining active learning strategies with self-labeling?

RQ2: What is the performance of the proposed method for datasets with a high number of learning examples?

RQ3: How does the initial training set size (the seed size) impact the performance?

RQ4: How does the accuracy of the model trained with seed impact the learning process of the proposed algorithm?

RQ5: Does the proposed algorithm allow for the better utilization of the computational budget?

Research questions

RQ1:Is there a benefit of combining active learning strategies with self-labeling?

RQ2: What is the performance of the proposed method for datasets with a high number of learning examples?

RQ3: How does the initial training set size (the seed size) impact the performance?

RQ4: How does the accuracy of the model trained with seed impact the learning process of the proposed algorithm?

RQ5: Does the proposed algorithm allow for the better utilization of the computational budget?

Research questions

RQ1:Is there a benefit of combining active learning strategies with self-labeling?

RQ2: What is the performance of the proposed method for datasets with a high number of learning examples?

RQ3: How does the initial training set size (the seed size) impact the performance?

RQ4: How does the accuracy of the model trained with seed impact the learning process of the proposed algorithm?

RQ5: Does the proposed algorithm allow for the better utilization of the computational budget?

Research questions

RQ1:Is there a benefit of combining active learning strategies with self-labeling?

RQ2: What is the performance of the proposed method for datasets with a high number of learning examples?

RQ3: How does the initial training set size (the seed size) impact the performance?

RQ4: How does the accuracy of the model trained with seed impact the learning process of the proposed algorithm?

RQ5: Does the proposed algorithm allow for the better utilization of the computational budget?

Datasets

dataset name	size	#class	#attributes	IR
adult	48842	2	14	3.153
bank marketing	45211	2	17	7.547
firewall	65478	3	12	2.929
chess	20902	15	40	22.919
nursery	12958	4	8	13.171
mushroom	8124	2	22	1.075
wine	4873	5	12	13.508
abalone	4098	11	8	21.542

Setup

- Metric: BAC (balanced accuracy)
- Protocol: All values of metrics were obtained with a separate test set and were averaged over runs with different random seeds.
- Code was implemented in Python with scikit-learn library.
- The codebase with the method and experiment implementations are available on github.

Baselines

- **random** - random selection of samples for query
- **fixed uncertainty** - selection of samples based on a static confidence thresholding
- **variable uncertainty** - modification of fixed uncertainty that adjusts confidence threshold based on the current size of the uncertainty region
- **classification margin** - a method that computes the difference in confidence between classes with two biggest supports
- **vote entropy** - queries are based on ensemble vote entropy
- **consensus entropy** - samples are selected based on the highest average prediction entropy
- **max disagreement** - computes KL-divergence between output class distribution and consensus distribution
- **min margin** - a method that selects samples based on minimum classification margin for all models in the ensemble

Comparision with SOTAs

- For small datasets: SLS2 rarely obtained the best score. However, the difference between the best-performing method and it was often close to or below 0.02.
- For big datasets: SL2S performed well, with either the best-BAC or close to the best. There is no clear performance pattern when we compare results across the varying budgets.
- SL2S could, obtain better performance with smaller seeds, and we could obtain the best BAC. This result indicates that SL2S does not depend heavily on the initial model performance and could be applied even if the number of labeled samples in the beginning, is small.

Comparision with SOTAs

- For small datasets: SLS2 rarely obtained the best score. However, the difference between the best-performing method and it was often close to or below 0.02.
- For big datasets: SL2S performed well, with either the best-BAC or close to the best. There is no clear performance pattern when we compare results across the varying budgets.
- SL2S could, obtain better performance with smaller seeds, and we could obtain the best BAC. This result indicates that SL2S does not depend heavily on the initial model performance and could be applied even if the number of labeled samples in the beginning, is small.

Comparision with SOTAs

- For small datasets: SLS2 rarely obtained the best score. However, the difference between the best-performing method and it was often close to or below 0.02.
- For big datasets: SL2S performed well, with either the best-BAC or close to the best. There is no clear performance pattern when we compare results across the varying budgets.
- SL2S could, obtain better performance with smaller seeds, and we could obtain the best BAC. This result indicates that SL2S does not depend heavily on the initial model performance and could be applied even if the number of labeled samples in the beginning, is small.

Ablation study

seed size	100	500	1000
base	0.4151±0.0220	0.4323±0.0230	0.4509±0.0131
-prior filter	0.3747±0.0192	0.4430±0.0193	0.4502±0.0231
-lambda reduction	0.3747±0.0192	0.4257±0.0159	0.4463±0.0226
-self-labeling	0.4174±0.0168	0.4421±0.0292	0.4461±0.0398
-bootstrapped training	0.3916±0.0211	0.4318±0.0217	0.4461±0.0135

- We found that the prior filter has a positive impact only in the case of smaller seed sizes.
- Reducing lambda after budget end provides gains in balanced accuracy for higher seed size.
- Removing bootstrapped training has a more significant impact when training with a smaller seed size.

Ablation study

seed size	100	500	1000
base	0.4151±0.0220	0.4323±0.0230	0.4509±0.0131
-prior filter	0.3747±0.0192	0.4430±0.0193	0.4502±0.0231
-lambda reduction	0.3747±0.0192	0.4257±0.0159	0.4463±0.0226
-self-labeling	0.4174±0.0168	0.4421±0.0292	0.4461±0.0398
-bootstrapped training	0.3916±0.0211	0.4318±0.0217	0.4461±0.0135

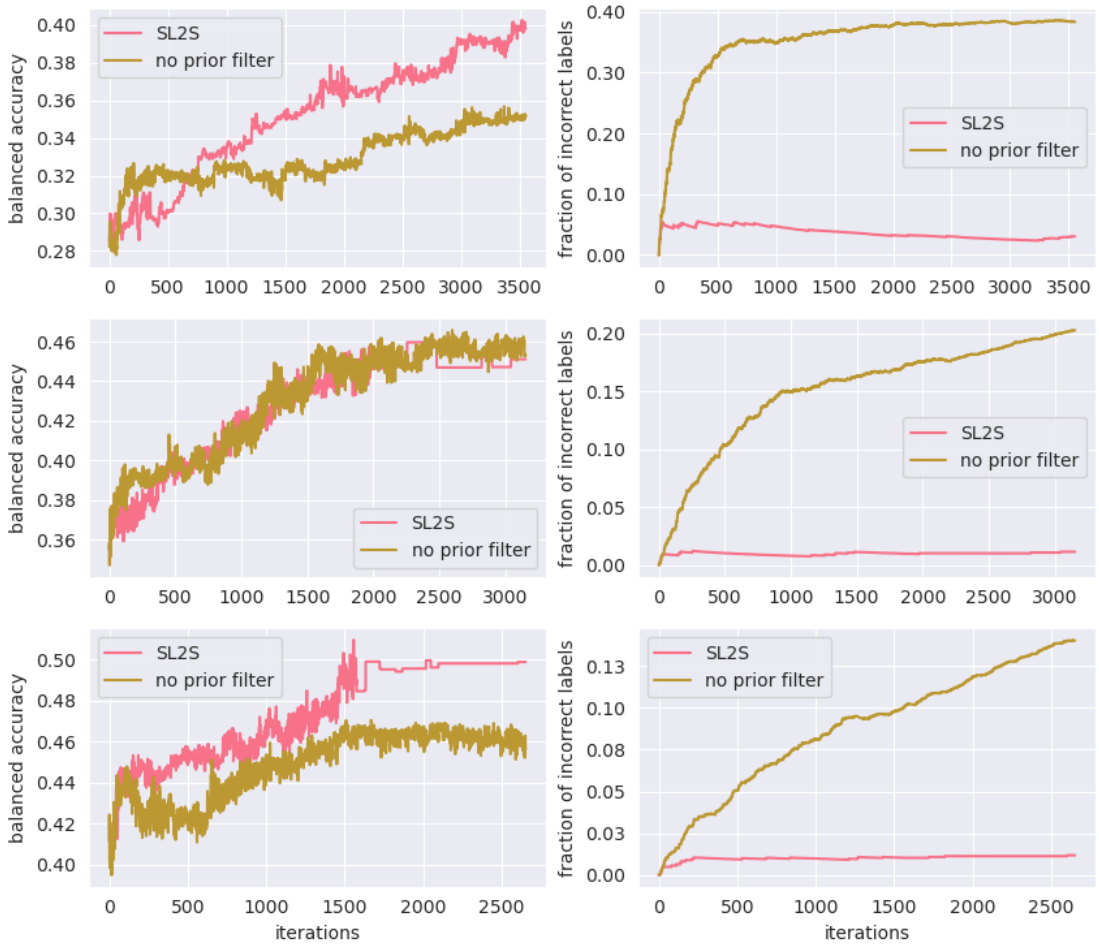
- We found that the prior filter has a positive impact only in the case of smaller seed sizes.
- Reducing lambda after budget end provides gains in balanced accuracy for higher seed size.
- Removing bootstrapped training has a more significant impact when training with a smaller seed size.

Ablation study

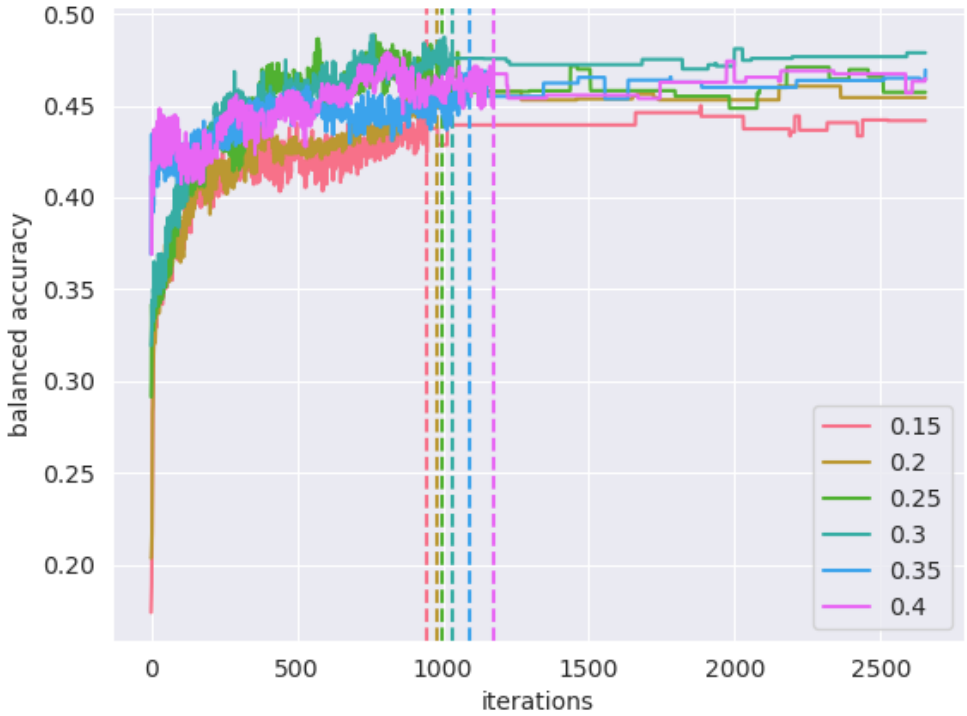
seed size	100	500	1000
base	0.4151±0.0220	0.4323±0.0230	0.4509±0.0131
-prior filter	0.3747±0.0192	0.4430±0.0193	0.4502±0.0231
-lambda reduction	0.3747±0.0192	0.4257±0.0159	0.4463±0.0226
-self-labeling	0.4174±0.0168	0.4421±0.0292	0.4461±0.0398
-bootstrapped training	0.3916±0.0211	0.4318±0.0217	0.4461±0.0135

- We found that the prior filter has a positive impact only in the case of smaller seed sizes.
- Reducing lambda after budget end provides gains in balanced accuracy for higher seed size.
- Removing bootstrapped training has a more significant impact when training with a smaller seed size.

Ablation study - extension



BAC with the corresponding fraction of samples with wrong labels in the training dataset over multiple iterations .



The relationship between the balanced accuracy of the initial model and overall experiment results. We evaluate models with initial accuracy equal to or exceeding 0.15, 0.2, 0.25, 0.3, 0.35, and 0.4, with a budget of 0.3.

Lessons learned

- SL2S works better for big datasets
- In the case of a small dataset, the performance is similar to other methods.
- The budget does not have a huge impact on the experiment results.
- A prior filter may not be the best method to address the imbalance issue in used datasets, but in the case of real datasets, the prior class distribution has higher importance. For this reason, alternative methods should be developed for dealing with imbalance when applying self-labeling to active learning scenarios.
- We showed that after the budget ends, the balanced accuracy roughly stays at the same level, and changes in the test accuracy do not occur frequently.

Lessons learned

- SL2S works better for big datasets
- In the case of a small dataset, the performance is similar to other methods.
- The budget does not have a huge impact on the experiment results.
- A prior filter may not be the best method to address the imbalance issue in used datasets, but in the case of real datasets, the prior class distribution has higher importance. For this reason, alternative methods should be developed for dealing with imbalance when applying self-labeling to active learning scenarios.
- We showed that after the budget ends, the balanced accuracy roughly stays at the same level, and changes in the test accuracy do not occur frequently.

Lessons learned

- SL2S works better for big datasets
- In the case of a small dataset, the performance is similar to other methods.
- The budget does not have a huge impact on the experiment results.
- A prior filter may not be the best method to address the imbalance issue in used datasets, but in the case of real datasets, the prior class distribution has higher importance. For this reason, alternative methods should be developed for dealing with imbalance when applying self-labeling to active learning scenarios.
- We showed that after the budget ends, the balanced accuracy roughly stays at the same level, and changes in the test accuracy do not occur frequently.

Lessons learned

- SL2S works better for big datasets
- In the case of a small dataset, the performance is similar to other methods.
- The budget does not have a huge impact on the experiment results.
- A prior filter may not be the best method to address the imbalance issue in used datasets, but in the case of real datasets, the prior class distribution has higher importance. For this reason, alternative methods should be developed for dealing with imbalance when applying self-labeling to active learning scenarios.
- We showed that after the budget ends, the balanced accuracy roughly stays at the same level, and changes in the test accuracy do not occur frequently.

Lessons learned

- SL2S works better for big datasets
- In the case of a small dataset, the performance is similar to other methods.
- The budget does not have a huge impact on the experiment results.
- A prior filter may not be the best method to address the imbalance issue in used datasets, but in the case of real datasets, the prior class distribution has higher importance. For this reason, alternative methods should be developed for dealing with imbalance when applying self-labeling to active learning scenarios.
- We showed that after the budget ends, the balanced accuracy roughly stays at the same level, and changes in the test accuracy do not occur frequently.

Outline

1. Motivations
2. Self Labeling Selective Sampling (SLS2)
3. Experiments
4. Conclusion

Conclusion

- We proposed SL2S a new active learning method that combines ensemble-based sample selection and self-labeling for selective sampling.
- Experiments with multiple baselines show that our algorithm offers comparable performance to other active learning algorithms for smaller datasets and better performance for bigger datasets.
- Further experiments show that our method works well when the initially labeled dataset is small or the initial model is poorly trained.

Conclusion

- We proposed SL2S a new active learning method that combines ensemble-based sample selection and self-labeling for selective sampling.
- Experiments with multiple baselines show that our algorithm offers comparable performance to other active learning algorithms for smaller datasets and better performance for bigger datasets.
- Further experiments show that our method works well when the initially labeled dataset is small or the initial model is poorly trained.

Conclusion

- We proposed SL2S a new active learning method that combines ensemble-based sample selection and self-labeling for selective sampling.
- Experiments with multiple baselines show that our algorithm offers comparable performance to other active learning algorithms for smaller datasets and better performance for bigger datasets.
- Further experiments show that our method works well when the initially labeled dataset is small or the initial model is poorly trained.

Aknowledgement

This work was supported by the CEUS-UNISONO programme, which has received funding from the National Science Centre, Poland under grant agreement No. 2020/02/Y/ST6/00037