Query Learning on the Semisupervised Margin

Miao Cheng

Email: *miaocheng@acm.org*
School of Computer Science
Guangxi Normal University
ICCIA'24, Haikou, China

August 4, 2024

Query Learning

- Query learning absorbs the **active** instances of classification into optimized approach.
- Ideally, the matched instance is able to **enhance** the learning ability of systems in iterative manner.
- Meanwhile, the subsets of data groups are to be **changed** along with active sampling.

Related Conceptions

Supervised Learning

All data are labeled with an identical category.

Semi-supervised Learning

Some labeled instances while a lot of unlabeled instances.

Related Conceptions

Incremental Learning

The patterns of coming data are sequentially collected, and **accumulative** ability is reached for classification of the specific data.

Online Learning

The learning ability of system is to be improved associated with **each** coming data, and it is to be predicated immediately for each data on hold.

Related Conceptions

Query Learning

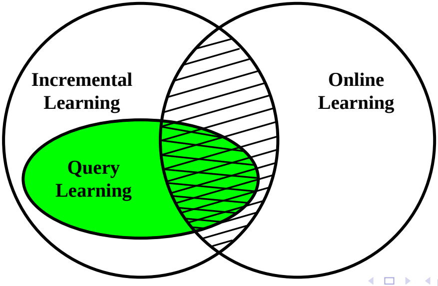
The learning performance is considered to be improved by **active data** of unlabeled set, which are **sampled** with respect to certain matched target.

Notice

- Those above methods rely on **the similar scenario** of data handling, and share **the common characteristics** of *sequential learning*.

Query Learning

- On behalf of query of active samples, query learning can be explained as an extension / variant of **incremental semi-supervised** learning.
- Query learning can be also enhanced by **discriminant models** of *online* learning.



Query Learning

Steps

- Given the learning models, calculate the maximum classification error of each unlabeled data.
- Select the most important instance with respect to the minimum error among all reference data.
- Absorb the selected data, and update the learning models.

Semisupervised Margin

- The respective classfication error associated with certain category is calculated accordingly,

$$err_i\left(\widetilde{x}, s\right) = \left\| {w_i}^T \widetilde{x} - s \right\|^2, \quad i = 1, 2, \cdots, c. \qquad (1)$$

- Thus, the classification errors of **each** category can be estimated by setting the **reference** label to be the **target** and **other labels** to be the **abnormal (deviate)** ones.

$$Err_j\left(\widetilde{x}\right) = \sum_{i=1, i \neq j}^{c} err_i(\widetilde{x}, -1)_i + err_j\left(\widetilde{x}, 1\right) \\ i = 1, 2, \cdots, c. \qquad (2)$$

Semisupervised Margin

- Learn the discriminant models for each data.
  - Regression
  - Deep learning
- Regressive learning

$$w_i = ((XX^T)^+ XL^T)_i, \qquad i = 1, 2, \cdots, c. \quad (3)$$

Query Learning

Steps

- Given the learning models, calculate the maximum classification error of each unlabeled data. (✔)
- Select the most important instance with respect to the minimum error among all reference data.
- Absorb the selected data, and update the learning models.

Update

Tips

- One instance is selected in a *single* circle.
- The leanring models are updated *iteratively*.

Deduction

- The online learning approaches can be adopted to update the discriminant models.

Update

- It can be reached by adopting existing online learning solutions with the estimated label of active sample.
- As a representative approach, relaxed online maximum margin (ROMM) is adopted to update the discriminant models.

Update

- The label of the current active sample is predicated again with the update approach.
- It is absorbed into the labeled subset, while removing from unlabeled subset.

Query Learning

Steps

- Given the learning models, calculate the maximum classification error of each unlabeled data. (✔)
- Select the most important instance with respect to the minimum error among all reference data.
- Absorb the selected data, and update the learning models. (✔)

Complexity

- The calculation of the initial w requires $O\left(d^3 + nd^2\right)$, and the prediction of errors requires $O\left(mcd\right)$ for all unlabeled data.
- The complexity of update is similar to that of the perceptron algorithm.
- The complexity of whole procedure relies on the query numbers.

Experiments

- Three benchmark data sets, namely Cifar 10, Food 101, Monkey.
- Besides the presented one, three other methods with calculation efficiency, i.e., Random, GAMBLE, UEER.

Experiments

- 100 images are randomly selected from each category of Cifar data set.
- The top 10 categories of Food are chosen.
- The *half* of each image data set are set to be the labeled data, while the rest images are used as unlabeled data.

Experiments

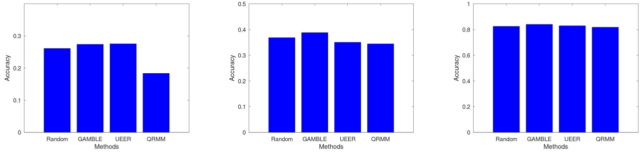


Figure: The obtained accuracy of different methods.

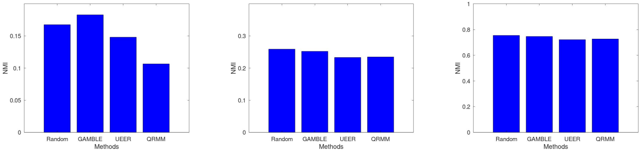


Figure: The obtained normalized mutual information of different methods.

Experiments

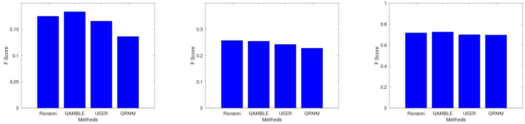


Figure: The obtained F scores of different methods.

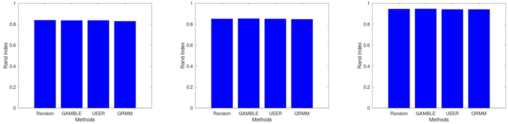


Figure: The obtained Rand index of different methods.

Conclusion

- An improved approach to active query is proposed, and the important instances are sampled with respect to the class-specific errors.
- Benefiting from online learning framework, the update can be done by adopting existing solutions.
- Experimental results validate the proposed approach with the optimistic results.

Thank You

@ ICCIA'24, Haikou, China