Explore some algorithms in active learning and semi-supervised learning for classification

Yubing Yao

1 Introduction

The dataset consists of a small number of labeled data-pairs of output and the features-$\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{l}$ and a large number of unlabeled data with the features only-$\{\boldsymbol{x}_j)\}_{j=l+1}^{l+u}$ when labels are scarce or expensive to obtain. Here I focus on classification problem where the outputs $\boldsymbol{y}$ are binary or multi-class. We explore some semi-supervised learning algorithms and active learning algoirthms for this type of mixture of labeled and unlabeled data. We compare those active learning and semi-supervised learning algorithms with closely related supervised learning counterparts for classification on typical simulated sets and three real data sets with either binary or multi-classes labels.

One typical setting of semi-supervised classification is to train a classifier f from both the labeled and unlabeled data, such that it is better than the supervised classifier trained on the labeled data alone under the assumption that trained data are well separated(2).

Active learning is a subfield of machine learning. A key hypothesis in active learning is that if the learning algorithm is allowed to choose the data from which it learns it will perform better with less training.

Supervised support vector machine(SVM) classifier is compared with semi-supervised learning SVM under the framework of Contrastive Pessimistic Likelihood(CPLE) on binary class labeled simulated and two real data sets. I explored how noisy levels by flipping well separated binary class labels would affect classification test accuracy scores of semi-supervised SVMs as compared to supervised SVM by cross validation. With increasing noisy levels of binary class labels, the classification test accuracy scores decrease in both semi-supervised SVM and supervised SVM. Both optimistic and pessimistic CPLE SVM methods are at least as comparable to supervised SVM classifier under different noisy levels of simulated binary class labeled data in terms of test accuracy scores. But in two binary labeled real data sets both optimistic and pessimistic CPLE SVM methods performs worse than supervised SVM classifier with similar size of labeled pairs and total size of mixture labeled pairs of features and unlabeled features to simulated datasets.

Also with increasing noisy levels in both simulated two-class or multi-class labeled datasets, the classification test accuracy scores decrease in several typical active learning algorithms such as Query by Committee, uncertainty sampling etc. More specifically in simulated two-class labeled datasets Query by Committee with KL divergence active learning algorithm is more robust to noisy class labels comparing with other active learning algorithms. And in simulated multi-class(4) labeled datasets it appears that uncertainty sampling with entropy active learning algorithm is more robust to noisy class labels comparing with other active learning algorithms.

2 Related Work

Semi-supervised CPLE likelihood based methods including pessimistic and optimistic CPLE have been compared under typical classifiers such as Linear Discriminant Analy-

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

sis (LDA), SVM etc to their corresponding supervised counterparts using some datasets in UCI machine learning data archive(4) and Scikit-learn library(examples in `semisup-learn libary:https://github.com/tmadl/semisup-learn`). A systematic investigation of the relationship between degrees of noisy labeled classes and Semi-supervised CPLE classifiers as compared to the supervised counterparts has been implemented here.

Also extensive literatures have compared various active learning algorithms(Query Strategies) and how the test accuracy scores increase with the increasing number of sampled unlabeled instances(1). Comparisons among typical active learning algorithms under varous noisy binary or multi-class labels in boh simulated and real mixtures of labeled and unlabeled datasets have been ran to check the robustness of these algorithms here.

3 Methodology

3.1 Algorithms in supervised classification

- Support Vector Machines classifier (SVM): a kernel based discriminative classifier formally defined by a separating hyperplane. The kernel used here is the (Gaussian) radial basis function kernel (RBF):

$$\exp(-\gamma|x-y|^2), \gamma > 0$$

- Logistic Regression: a probabilistic discriminative classifier, with the binary label-$Y = 0, 1$, then

$$P(Y = 1|X) = \frac{1}{1 + \exp(-(W^T X + b))}$$

  With multi-class labels, the algorithm applies one verus rest mechanism in default parameter setup in Scikit-learn LogisticRegressoin function.

These two supervised classification algorithms have been implemented in semi-supervised learning and active learning respectively.

3.2 Algorithms in semi-supervised classification

Likelihood-based Classifier-Contrastive Pessimistic Likelihood (CPLE): a general way to perform semi-supervised parameter estimation for likelihood-based classifiers(4) and it has been generalized to other types of classifier(arbitrary scikit-learn classifier) such as kernel based Support Vector Machines in Python library-`semisup-learn:https://github.com/tmadl/semisup-learn`. This method implemented in `semisup-learn` library not only assumes that the true labels of the unlabeled instances are as adversarial to the likelihood as possible, and also tries to increase the likelihood on the labeled examples.

- Pessimistic CPLE: the label hypotheses for the unlabeled instances should be pessimistic (i.e. minimize log likelihood)
- Optimistic CPLE: the label hypotheses for the unlabeled instances should be optimistic (i.e. maximize log likelihood)

3.3 Pool-Based Sampling Algorithms in active learning for classification

Active learning algorithms ask queries in the form of unlabeled instances to be labeled by an `oracle`(e.g. a human annotator) where the below active learning algorithms apply different query algorithms:

- Random Sampling: Randomly sample the instances in unlabeled set to label and pool together with the labeled pairs to fit the classification model. This method is used as a baseline comparing with other active learning algorithms.
- Uncertainty Sampling: The query algorithm in active learning apply uncertainty sampling method. In this algorithm, an active learner queries the samples about which it is least certain how to label. This algorithm was used for the experimental and real data set with

multi-classes labels. The three uncertain criteria in uncertainty sampling algorithms have been used:
1. Query the samples whose prediction are `least confident`:

$$x^*_{lc} = \underset{x}{\operatorname{argmax}}(1 - P_\theta(\hat{y}|x))$$

where $\hat{y} = argmax_y P_\theta(y|x)$, or the class label with the highest posterior probability under the model θ.
2. Another more general uncertainty sampling strategy is to use `entropy` as the uncertainty measure:

$$x^*_H = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

where y_i ranges over all possible labelings.
3. A different multi-class uncertainty sampling criterion is to sample the instance with `smallest margin`:

$$x^*_{sm} = \underset{x}{\operatorname{argmin}}(P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x))$$

where $\hat{y}_1, \hat{y}_2$ are the first and second most probable class labels under the model respectively.

- Query By Committee (QBC) algorithm: A committee of models are trained on labeled data, but representing the competing hypotheses. Each committee member is then allowed to vote on the labeling of the query candidates, The most informative query is considered to be the instance they most disgreee about. Two main approaches are used to measure the level of disagreement.
1. The first disagreement measure is `vote entropy`

$$x^*_{vote} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

where y_i ranges over all possible labelings and $V(y_i)$ is the number of votes that a label receives from among the committee members' predictions, and C is the committee size.
2. The second disagreement measure is average `Kullback-leiber(KL) divergence`:

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} - \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}}||P_C)$$

where $D(P_{\theta^{(c)}}||P_C) = \sum_i P_{\theta^{(c)}}(y_i|x) \log \frac{P_{\theta^{(c)}}(y_i|x)}{P_C(y_i|x)}$, and $\theta^{(c)}$ represents a particular model in the committee and C represents the committee as a whole.

- Expected Error Reduction (EER): EER active learning algorithm concentrates on how much the generalization error is likely to reduced after quering the samples. The rough idea it to estimate the expected future error of a model and query the instance with minimal expected future error. The objective here is to reduce the expected total number of incorrect predictions through minimizing the expected log-loss:

$$x^*_{log} = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \left(- \sum_{u=1}^{U} \sum_j P_{\theta^{+<x,y_i>}}(y_j|x^{(u)}) \log P_{\theta^{+<x,y_i>}}(y_j|x^{(u)}) \right)$$

- Density Weighted Uncertainty Sampling (DWUS): This active learning algorithm firstly estimates the distribution(density) of the entire input spaces through k-means clustering method to propogate label information to instances in the same cluster assumng clustering structures of input space. This method is less prone to querying outliers than other query strategies such as QBC, uncertainty sampling.

3.4 Cross validation

The fully labeled data sets have been randomly split into train (appriximately 2/3 of the whole data set)and test sets (1/3).The train sets have been transformed to include a small size of labeled samples and others unlabeled. The semi-supervised and active learning algorithms will be fit to train sets with mixtures of labeled and unlabeled instances and the prediction of class labels and accuracy scores will be carried out on test sets. The above procedure will be repeated 10 times to generate the avarage test accuracy scores of corresponding algoirthms.

4 Data sets

4.1 Simulated Datasets

The simulated classification fully labeled datasets are generated through `make_classification` in sklearn.datasets with size 500 and the number of features-20 in both binary and multi-class(4) labels. The noisy levels of class labels are through the percentages of flipping well separated labels-0.01,0.1,0.3,0.5 in `make_classification`. The fully labeled sets need to be transformed as mixture of labeled (size-50, 150) and unlabeled components in train sets to fit semi-supervised learning and active learning classification models.

4.2 Real datasets

Two real datasets-Wisconsin Diagnostic Breast Cancer (wdbc)(5) and breast-cancer-wisconsindata(6) with binary class labels from UC Irvine Machine Learning Repository are used(`http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`). The detailed description of class labels and the features and other information in these two datasets are from the documents of web links-`http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names` and `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names` respectively. The two real datasets will be artificially transformed to mixture of labeled and unlabeled sets by similar stratety in simulated datasets.

Another real dataset with multi-class labeling addressed in project proposal is more than 10000 observations and 561 features with 6 class labels. The running time of active learning algorithms is too long and is at least more than 12 hours even after randomly sampling 1000 observations from the whole data set. Thus this dataset is no long used in final project report.

5 Experiments and Results

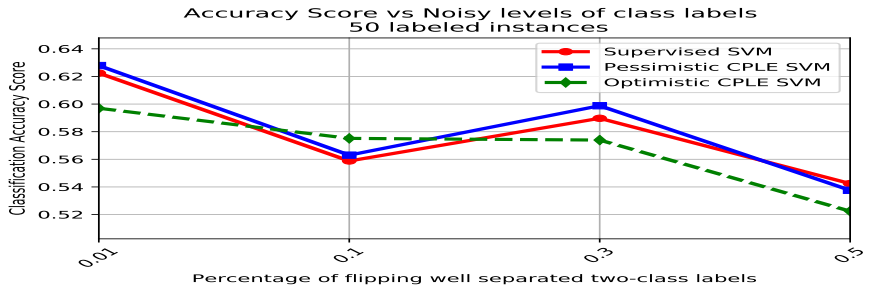
5.1 Results using semi-supervised learning algorithms

5.1.1 Results from Simulated Data sets

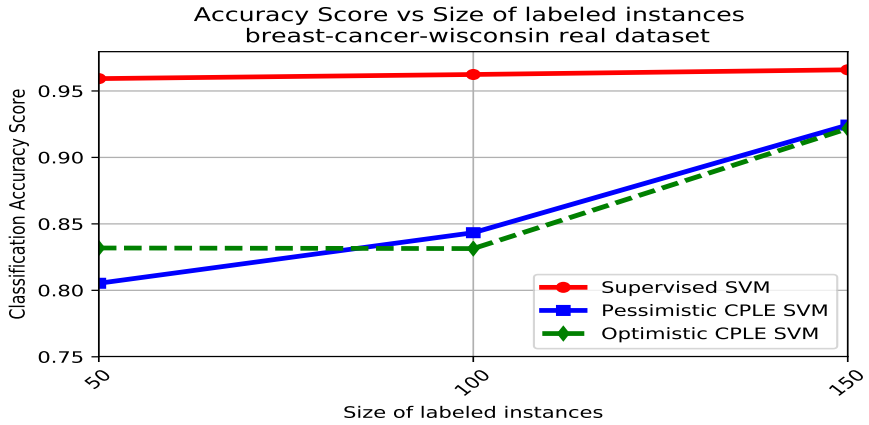
I explored the robustness of semi-supervised learning algorithms-pessimistic and optimistic CPLE SVMs with supervised SVM on simulated through different percentages of flipping well separated two-class labels-0.01, 0.1,0.3 and 0.5. The average test accruacy scores of four noisy levels of two-class labels with size of labeled instances 50 are shown below:
Overally the average test accuracy scores from three methods-supervised SVM, semi-supervised pessimistic CPLE SVM, optimistic CPLE SVM decrease with incrasing noisy levels of two-class labels in this experimental classification scenario. Test accuracy scores from semi-supervised pessimistic CPLE SVM are consistently slightly better or very close to the correponding ones from supervised SVM under different noisy levels, while the results from optimistic CPLE SVM are relatively more different from the corresponding ones from other two methods. Also results with these three algorithms on 150 labeled instances show similar tendency.

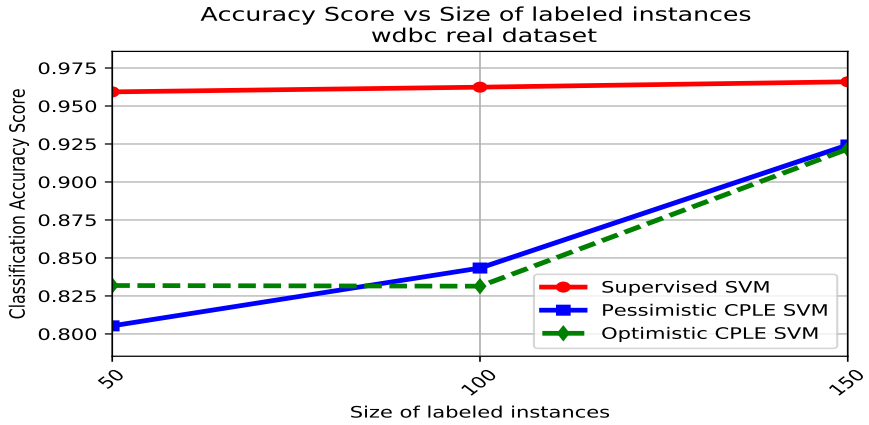
5.1.2 Results from two real Data sets

Apply cross-validation method and transform the fully labeled train set to mixture of labeled and unlabeled components with size of labeled intances-50,100,150 on two real datasets-Wisconsin



Diagnostic Breast Cancer (wdbc) and breast-cancer-wisconsindata, repeat the above experiment 10 times and average the 10 classification test accuracy scores. The average test accuracy scores of three methods-supervised SVM, semi-supervised pessimistic CPLE SVM, optimistic CPLE SVM under different sizes of labeled instances on two real datasets are shown below respectively:





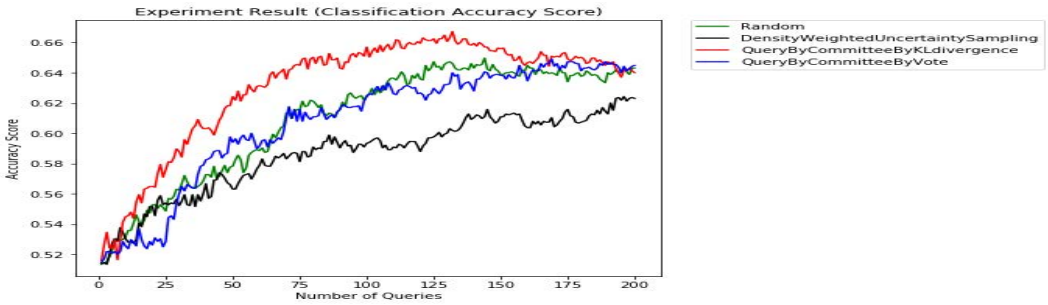
Supervised SVM classifiers on these two real datasets have consistently better performance in terms of test accuracy scores than other two semi-supervised CPLE SVM methods on both real datasets. With increasing size of labeled instances, test accuracy scores from both semi-supervised CPLE SVM are closer to the one from supervised SVM in both real datasets

5.2 Results from active learning algorithms

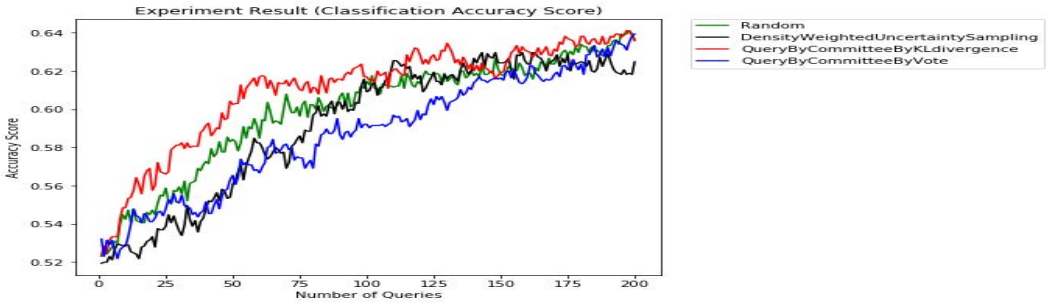
5.2.1 Simulated two-class and four-class datasets

I explored the robustness of typical active learning algorithms for both binary and multiple-class labels in `libact` Python library under different degrees of noisy class labels through percentages of flipping well separated class labels-0.01, 0.1,0.3 and 0.5 in 500 size and 20 features simulated datasets, where logistic regression is used as classification model. The transformed mixture of labeled and unlabeled components contains 50 labeled instances.
The average test accruacy scores of four noisy levels of two-class labels with ideal labeling on 1 to 200 queried unlabeled samples sequentially are shown below with four active learning algorithms-random sampling, Density Weighted Uncertainty Sampling, Query by Committee by vote and Query by Committee by KL divergence. With moderate noisy two-class labels Query by

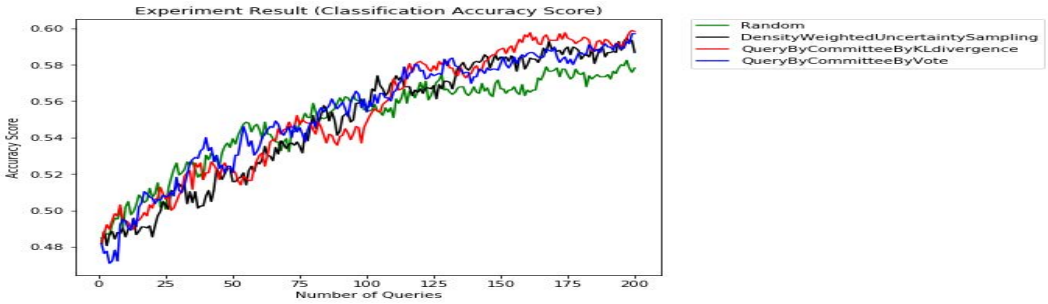
Percentage of flipping well separated two-class labels:0.01



Percentage of flipping well separated two-class labels:0.1

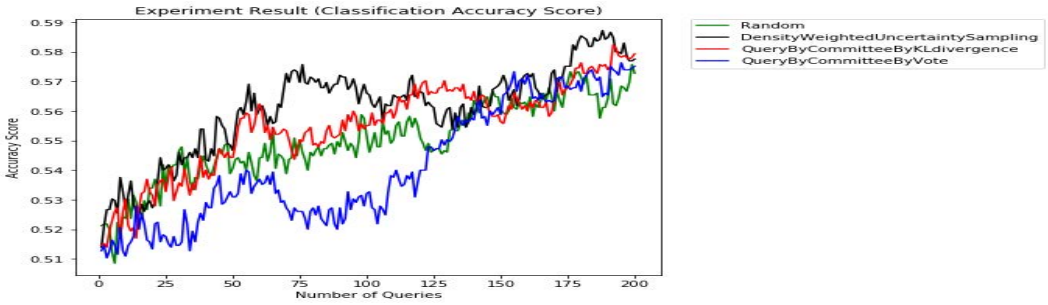


Percentage of flipping well separated two-class labels:0.3



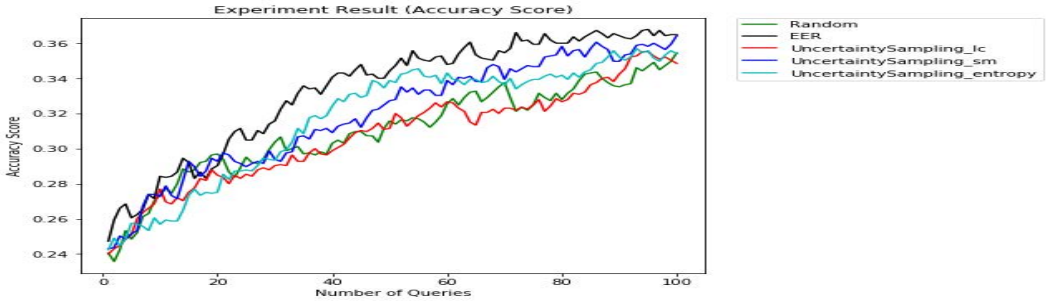
Committee by KL divergence has better performance in test accuracy score than other three active learning across different numbers of queries from 1 to 200. With increasing noisy levels of two-class labels all four active learning algorithms all have a decreasing tendency of test accuracy scores under same numbers of queries.

Percentage of flipping well separated two-class labels:0.5

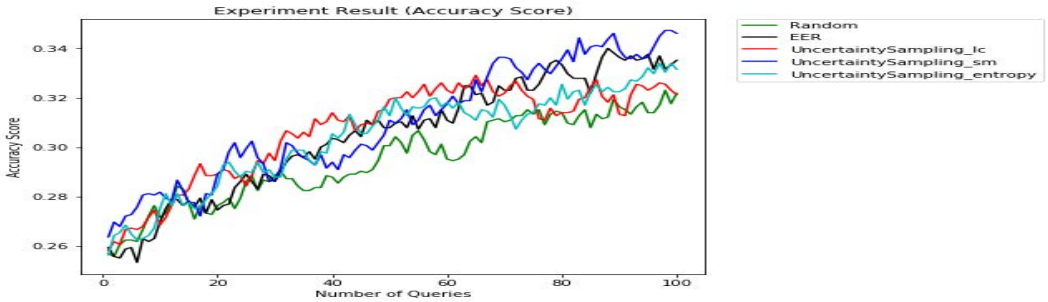


The average test accruacy scores of four noisy levels of multi-class(4) labels with ideal labeling on 1 to 200 queried unlabeled samples sequentially are shown below with five active learning algorithms-Random sampling, Expected Error Reduction (EER), Uncertainty Sampling with least confidence(ls), Uncertainty Sampling with smallest margin(sm) and Uncertainty Sampling with entropy: The same decreasing tendency in terms of classification test accuracy score among all five

Percentage of flipping well separated multi-class(4) labels:0.01



Percentage of flipping well separated multi-class(4) labels:0.1

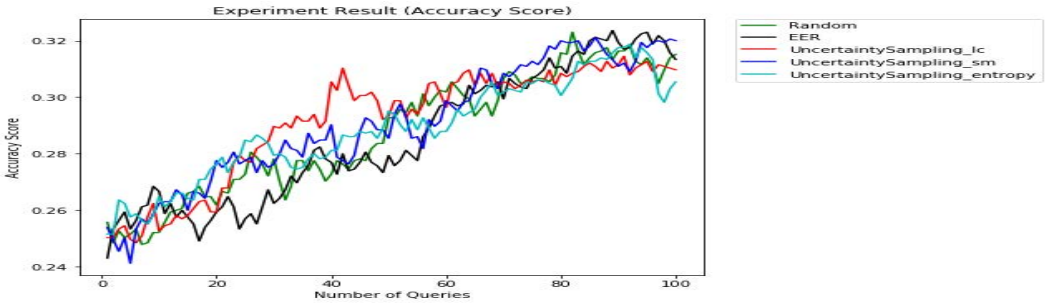


active learning algorithms exists with increasing noisy levels of multi-class(4) labels across the same number of queries. When multi-class labels are well separated, EER active learning algorithm, as expected, has better performance in test accuracy scores than other four active learning algorithms as the number of queries is larger than 20, however, this superiority of better classification accuracy in EER active learning algorithm didn't exist when the noisy level of multi-class labels becomes high as shown in the results where percentages of flipping well separated multi-class labels is 0.1,0.3,0.5.

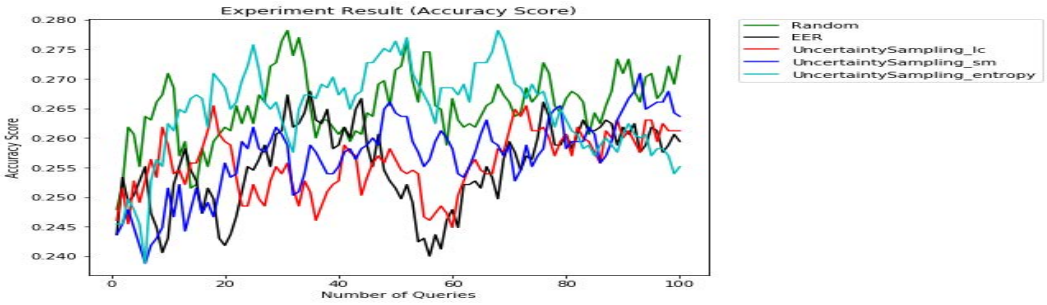
5.2.2 Performance of active learning algorithms in two real datasets

Consistent with the finding from simulated datasets with two-class labels, Query by Committee by KL divergence active learning algorithm has better performance in test accuracy scores than other active

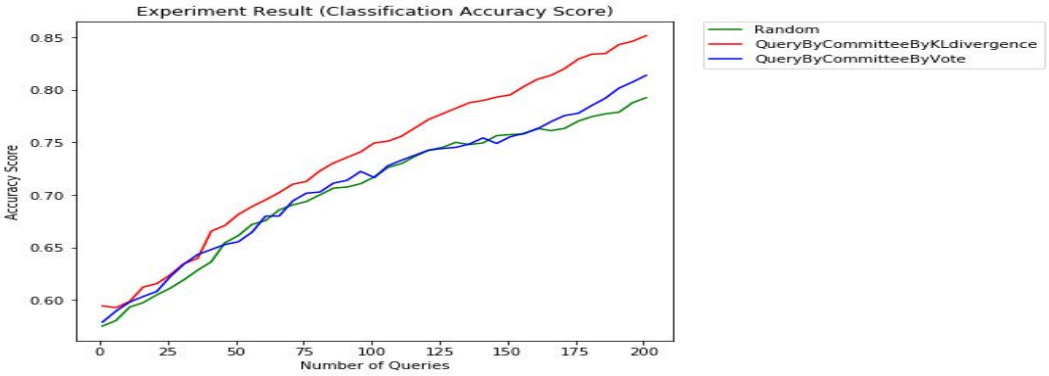
Percentage of flipping well separated multi-class(4) labels:0.3



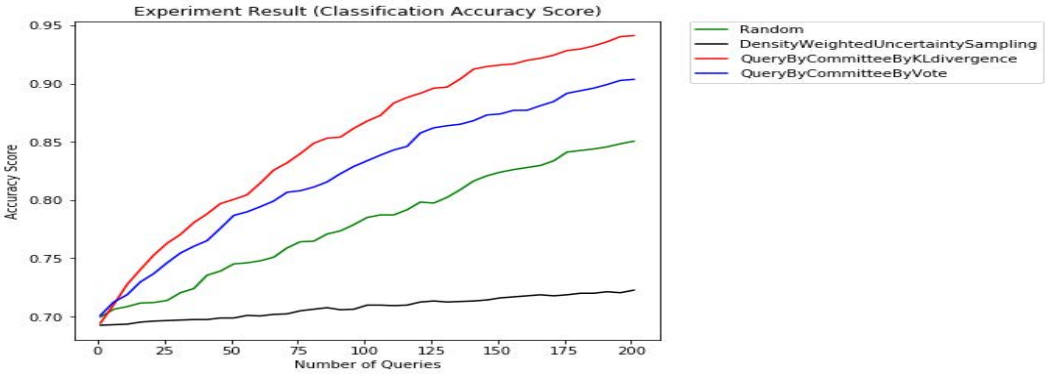
Percentage of flipping well separated multi-class(4) labels:0.5



Comparison of active learning algorithms on breast-cancer-wisconsin data with two-class labels



Comparison of active learning algorithms on wdbc data with two-class labels



learning algorithms across numbers of queries from 1 to 200 on both real datasets with two-class labels.

6 Discussion and Conclusions

Robustness of some semi-supervised learning and active learning algorithms in terms of noisy levels of class labels has been systematically explored, where the mixture of labeled and unlabeled components data is either with either two-class or multi-class labels. Among semi-supervised learning algorithms and active learning algorithms most of them are vulnerable to increasing noisy levels of class labels, with decreasing classification accuracy when the class labels are not well separated. Those results suggsts those semi-supervised learning and active learning algorithms may not have a mechanism to deal with very noisy class labels.

For semi-supervised learning algorithms, pessimistic CPLE classifier is at least as comparable to supervised counterpart in terms of classification accuracy. This result is based on likelihood ((4)) and thus any generative classifiers would have similar behavior. In semiup-learn library, this CPLE framework has been extended to arbitrary sklearn classifier such as K-nearest neighbor(KNN) and decision tree. Here we explored classificaton accuracy of pessimistic and optimistic CPLE SVMs comparing to supervised SVM in both simulated and real datasets. The classification accuracy results comparing pessimistic CPLE SVM to supervised SVM are not consitent in simulated and real datasets suggesting that results of likelihood based CPLE framework may not be well generalized to arbitrary classifier in terms of comparable classification accuracy to supervised counterpart.

For active learning algorithms, QBC with KL divergence for two-class labels, has superior classification efficiency for both moderate noisy simulated datasets and also two real datasets with logistic regression; similar superior tendency of EER to other active learning algorithms for multi-class labels exists in well separated simulated datasets, which are consistent with some results mentioned in the active learning literature survey ((1)).

References

[1] Burr Settles. Active Learning Literature Survey. *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison, 2009.

[2] Xiaojin Zhu and Andrew B. Introduction to Semi-Supervised Learning *Goldberg Synthesis Lectures on Artificial Intelligence and Machine Learning* , 2009.

[3] Yao-Yuan Yang, Yu-An Chung, Shao-Chuan Lee, Tung-En Wu, Hsuan-Tien Lin Libact: Pool-based Active Learning in Python `https://github.com/ntucllab/libact`, 2015.

[4] Marco Loog Contrastive Pessimistic Likelihood Estimation for Semi-Supervised Classification *arXiv:1503.00269*, 2015.

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64:304-310, 1989.

[6] W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *International Symposium on Electronic Imaging: Science and Technology*, 1905: 861-870, 1993.