Homework#1

(CSE 584)

Paper 1: "Active Learning by Learning"

Authors: Wei-Ning Hsu, Hsuan-Tien Lin

1. What problem does this paper try to solve, i.e., its motivation?

This paper addresses the challenge in the context of pool-based active learning whereby identifying the most suitable method for selecting instances to label is normally done by guesswork by humans, and may not be consistent across different datasets. Current approaches to active learning are centered on philosophies that are designed and developed by humans which may not be effective in all cases.

2. How does it solve the problem?

The paper suggests the new approach which is named Active Learning by Learning (ALBL) and compared to the multi-armed bandit problem.

- ALBL divides each candidate active learning algorithm into so called “bandit machine” and employs the bandit algorithm of type EXP4. P algorithm to update the sampling weights of these above-mentioned algorithms to optimize it on the given dataset.
- To establish the correspondence between the active learning goal and the multi-armed bandit problem, ALBL uses a reward function referred to as IMPORTANCE-WEIGHTED-ACCURACY (IW-ACC).
- This form of reward function offers an accurate evaluation, i.e. unbiased estimate, of the performance of the test accuracy unlike the training accuracy which is rather biased.
- Moreover, to ensure an accurate estimation of test accuracy using IW-ACC, ALBL employs a RANDOM sampling strategy through which the model may request labelled examples again.

3. A list of novelties/contributions?

- ALBL Framework: The paper proposed an adaptive active learning approach based on the multi-armed bandit problem to dynamically select and interpolate several active learning strategies.

- IW-ACC Reward Function: It suggests the employment of IW-ACC as reward function that is used to get an unbiased estimator of the test accuracy, and, therefore, make the multi-armed bandit optimization correspond directly to the measure of performance being optimized.
- Incorporation of RANDOM Strategy: Besides, the presence of a RANDOM strategy helps in unbiased estimation of test accuracy and acts as a competitive baseline strategy as well as a last resort in case the other strategies do not work.
- Empirical Validation: In this paper, the authors have provided sufficient empirical results to show that ALBL is better than single active learning methods, combination of two fixed methods and has better result than the only adaptive procedure COMB in datasets they used.

4. What do you think are the downsides of the work?

- Computational Complexity: ALBL employs the concept of using several active learning algorithms, as well as dynamic weight updates which may, therefore, add to the algorithm's complexity compared to utilizing a single fixed active learning approach.
- Dependence on the Quality of Base Algorithms: The performance of the proposed method of ALBL depends on the efficiency of the candidate active learning algorithms given above. As with most approaches that rely on base algorithms, if the base algorithms are not ideal for the dataset, then the performance of ALBL might be tame.
- Limited Exploration in Later Stages: As ALBL evolves, it may transform into specific active learning algorithm based on the estimated reward. This may lead to a neglect of the other algorithms that may be more efficient as the program develops to the next stages.

Paper 2: "Learning Active Learning from Data"

Authors: Ksenia Konyushkova, Sznitman Raphael, Pascal Fua

1. What problem does this paper try to solve, i.e., its motivation?

The paper aims to overcome an important difficulty of collecting labelled data for machine learning where labelling in certain domains takes considerable time and hiring experts is not affordable. Current AL strategies, though meant to reduce the amount of annotation done, do not portray standard results when performed on the same task or any different domain.

2. How does it solve the problem?

The paper presents a new concept in the selection of the samples to be annotated that is called the Learning Active Learning (LAL). This approach entails learning a regression model (They used Random Forest in their experiments), for estimating the degree of generalization error reduction which adding a given data instance can bring, using uncertainty sampling (US), the most frequently-used AL heuristic.

3. A list of novelties/contributions?

- Data-Driven Strategy: Unlike other approaches that rely on pre-specified heuristics, LAL or Learning of AL strategies is learned directly from data.
- Continuum of Strategies: LAL takes all realizations of AL strategies as a spectrum which is continuous while meta-AL approaches are based on combining some heuristics.
- Transferability: Domain adaptability can be achieved in LAL since strategies learned can be transferred to the target domain so that labelled data is not much of a necessity in a new domain.
- Two Implementations: To overcome this problem, two alternatives of LAL are considered in this paper: LALINDEPENDENT which adds data points randomly and LALITERATIVE which tries to mimic the iterative AL process to correct selection bias.

4. What do you think are the downsides of the work?

- Computational Cost: Although the online phase of LAL is concise, the offline procedure of obtaining datasets to train the regression model may require significant amounts of runs, particularly in the iterative variant.
- Limited Scope: The paper specialises in binary classification tasks and never directly speak about multi-class cases or batch-mode AL, however, both can be viewed as additional potential developments.
- Dependence on Representative Data: Since, LAL depends on the regression analysis, one important factor is that an appropriate sample set, synthetic or from another domain, should be available for the training of the model.

Paper 3: "Pool-based Active Learning based on Incremental Decision Tree"

Authors: Shuo Wang, Jian-Jian Wang, Xiang-Hui Gao, Xue-Zheng Wang

1. What problem does this paper try to solve, i.e., its motivation?

The paper aims to respond to the difficulty of constructing effective classifiers in the cases where the amount of labelled data is limited and acquires a high price. This is typical in most machine learning tasks where we have a lot of data available but it is expensive to label them.

2. How does it solve the problem?

The problem can be resolved by the use of an active learning algorithm that can be referred to as the Sample Selection Based on the Incremental Decision Tree or abbreviated as IDTSS as presented in this paper.

- IDTSS is under the group of pool-based active learning where it randomly has a pool of unlabelled data from which it samples on the best samples, i.e. most informative, that should be labelled by an expert. The algorithm has adopted the incremental decision tree learning approach that enables the decision tree classifier's optimization each time new labelled samples are added.
- The essence of IDTSS is the sample selection phase. While it avoids giving a direct measure of the uncertainty of an unlabelled sample, the design of IDTSS is to select samples based on the guidance of the whole alphabet of unlabelled samples. In the process of update, the algorithm first introduces the unlabelled sample to the training process with both positive and negative labels, and build two almost similar decision trees. The final step is to compare the classification of the rest of the unlabelled samples with the one defined by these two trees. The unlabelled sample that leads to the greatest disagreement is considered as the most useful and thus labelled for further use.

3. A list of novelties/contributions?

- New Sample Selection Mechanism: The most important novelty is in the sample selection strategy based on calculating the disagreement in the pool of unlabelled data due to hypothetical labels. It is different from other approaches, which are based on the kind of uncertainty of individual samples setting.

- Integration of Incremental Decision Trees: The roles of the incremental decision tree learning in the active learning framework are well presented in the paper and it is possible to include the new labelled data to update the classifier on the status.
- Empirical Evaluation: The authors perform experiments on several datasets collected from the UCI Machine Learning Repository that shows the efficiency of the proposed IDTSS algorithm in regards to other state-of-the-art methods of active learning especially in terms of its ability to achieve higher levels of accuracy using a limited number of labelled samples.

4. What do you think are the downsides of the work?

- Computational Cost: The integration technique involves performing a new decision tree training for each of the iteratively selected but unlabelled sample. This process could have been time consuming or may have required a lot of computations especially in cases where large number of data sets are involved or when the tree is deep or has many branches.
- Sensitivity to Initial Training Set: There may be the issue of dependency of the algorithm to the initial training set used in creating the initial decision tree. One of the major drawbacks is that the training set should include a small or not very representative collection of samples, which influences the possibilities of using informative samples for learning.
- Limited Exploration of Different Datasets: The experiments are performed only with a few datasets taken from the same repository. Evaluation on a broader scale, using a host of datasets including high-dimensional or composed feature space is needed to ascertain the general validity of the work that has been proposed.