Augmented Query Strategies for Active Learning in Stream Data Mining

Mustafa Amir Faisal[1], Zeyar Aung[2⋆], Wei Lee Woon[2], and Davor Svetinovic[2]

[1] Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA.
[2] Institute Center for Smart and Sustainable Systems (iSmart), Masdar Institute of Science and Technology, Abu Dhabi 54224, UAE.
mustafa.faisal@utdallas.edu,{zaung,wwoon,dsvetinovic}@masdar.ac.ae

Abstract. Active learning is used in situations where the amount of unlabeled data is abundant but it is costly to manually label the data. So, depending on our available budget, from all unlabeled instances we are to select only a subset of them to ask the oracle for manual labeling. Thus, the query strategy, i.e., how relevant instances are selected to be sent to the oracle, plays an important role in active learning. Though active learning is a very established research area, only a few research works have been done on it in the context of stream data mining. Active learning for stream data is more challenging than for static data because the repetition of queries is not feasible as revisiting the data is almost impossible. In this paper, we propose two augmented query strategies for active learning in stream data mining, namely, Margin Sampling with Variable Uncertainty (MSVU) and Entropy Sampling with Uncertainty using Randomization (ESUR). These two strategies are derived and improved from the existing methods of Variable Uncertainty (VU) and Uncertainty using Randomization (UR) respectively. We evaluate the effectiveness of our proposed MSVU and ESUR strategies by comparing them against the original VU and UR on 6 different datasets using two base classifiers: Leveraging Bagging (LB) and Single Classifier Drift (SCD). Experimental results show that our proposed strategies offer promising outcomes for various datasets and detecting concept drift in the data.

Keywords: Stream data mining, Active learning, Query strategy.

1 Introduction

Active learning has been a popular area of research since the 1990s. It is very useful in machine learning applications in which the amount of unlabeled data is abundant but manually labeling the data is costly. An example is spam filtering where it would be very difficult to manually label all instances in a training set. Thus, we need to select right instances to ask the oracle for manual labeling because of the limited labeling "budget". In the case of a large mail server only a relatively small subset, say hundreds to thousands (from amongst millions of

⋆ Corresponding author.

available emails) need to be manually labeled (it is assumed that our oracle, i.e., the manual labeling process, is not noisy and always provides correct labels.) Then, these selected instances and their manually assigned labels can be used to build an automatic classifier. The main motivation of active learning is to use very small amounts of manually labeled data to train classifiers while at the same time keeping accuracy high. Therefore, the query strategy, i.e., the procedure for selecting instances to be sent to the oracle, plays an important role in active learning.

Though active learning is a very established area of research, very little research have been done on it in the context of stream data mining. Active learning in stream data mining imposes more challenges than active learning in static data mining. In stream mining, repetition of a query is not feasible as revisiting data is almost impossible. Moreover, historical data cannot be stored because of limited functional memory. Hence, with a limited amount of labeled data, maintaining high accuracy is a crucial challenge.

In this paper, the focus is on active learning query strategies with stream data. Instead of pool-based sampling we assume that data cannot be buffered and a decision should be made for each data instance. Thus, we are interested in stream-based selective sampling using different query strategies. The motivation for this is to use active learning for stream data mining in small devices like active RFID tags [1], wireless sensor nodes [2], and smart meters [3], etc. in which the amount of memory is very limited. In addition, we also consider evolving nature of data where concept drift can happen.

The main objective of this research is to investigate the existing query strategies and to develop new ones that outperform the existing approaches — particularly in the context of stream data mining. Consequently, our main contributions in this paper are that: (1) We have proposed two new query strategies for stream data mining, namely MSVU and ESUR, by enhancing the existing state-of-the-art ones, namely VU and UR respectively. We have shown that the proposed MSVU and ESUR strategies outperform their original counterparts in a majority of test cases. (2) We have made some important observations regarding various query strategies that their performances vary greatly depending on the base classifier or the change detection technique used.

2 Relevant Background

Research on active learning with stream data is comparatively new while more work has been done with static data. Only a brief overview is provided in this section. For more detailed information, readers are referred to [4], which provides a comprehensive survey of existing query strategies for active learning in both static and stream settings.

The Random Strategy is a very basic one in which the learner selects random instances which are then presented to the oracle for labeling [5]. Every incoming instance is presented to the oracle with probability β, which is the pre-defined budget.

The Uncertainty Sampling Strategy is a general strategy first introduced by [6] where the learner asks the oracle about the instances about which it is the

least certain. There are a number of subcategories of this strategy, which are described below.

2.1 Least Confidence

- Fixed Uncertainty: In this strategy for online learning, as described in [5], instances for which the certainty is below a fixed threshold are flagged for labeling, where certainty is based on the posterior probability estimates provided by a classifier or learner. However, this is not practical in stream data mining, where data evolves quickly and with constant concept drift.
- Variable Uncertainty (VU): To overcome the limitations of the fixed uncertainty strategy, Žliobaitė *et al.* [5] introduced a variable threshold which adapts to the changing characteristics of the data.
- Uncertainty using Randomization (UR): In data stream variation or concept drift can occur anywhere in the input space. However, uncertainty strategy labels the instances that are near to the decision boundary. To mitigate this problem, the labeling threshold is randomized by multiplying by a normally distributed random variable that is within $\mathcal{N}(1, \sigma^2)$ [5].

In Section 3, we propose two augmented strategies based on VU and UR respectively and compare the performances of the augmented strategies with those of the original ones.

2.2 Margin Sampling

In the Least Confidence strategy, the most possible label is considered. This may lead to information about the remaining label distribution being ignored. An attempt has been made to correct this shortcoming using a different multi-class uncertainty sampling variant named Margin Sampling [7]. The definition of margin sampling is $\boldsymbol{X}_C = \arg\min_{D_t} P_C(\hat{y}_1|D_t) - P_C(\hat{y}_2|D_t)$, where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most possible class labels respectively under model C, and D_t is an incoming instance at time t.

In this paper, we amalgamate the idea of margin sampling and the Variable Uncertainty (VU) strategy [5] to come up with a better method of Margin Sampling with Variable Uncertainty (MSVU) as described below in Section 3.

2.3 Entropy

Entropy [8], an information-theoretic measure, is an uncertainty gauge presenting the amount of information required to encode a distribution. The definition of entropy is: $X_H^* = \arg\max_x (-\sum_i P_C(y_i|x) \log P_C(y_i|x))$, where y_i ranges over all possible labelings under model C. Entropy is usually regarded as a measure of uncertainty or impurity in machine learning. The entropy-based approach generalizes well to probabilistic multi-label classifiers and probabilistic models for more complex data like sequences [4].

In this paper, we enhance the Uncertainty using Randomization (UR) strategy [5] by incorporating sequence entropy in order to develop a better strategy named Entropy Sampling with Uncertainty using Randomization (ESUR) as descried below in Section 3.

3 Proposed Query Strategies

Let us consider $D_1, D_2, D_3, \ldots, D_t, \ldots$ as a data stream, where D_t as an instance at time t. The budget β indicates that in incoming data, $\beta\%$ of data are expected to be labeled by the oracle. The assumption is that labeling cost is same for every instance. Each query strategy takes an instance D_t, budget β, and other necessary parameters and decides whether to ask for labeling or not. After getting the label, the classifier is trained with this instance until budget, β is not exhausted.

In [5], the authors proposed two basic strategies with least confidence for uncertainty sampling, namely, Variable Uncertainty (VU) and Uncertainty using Randomization (UR). In our work, we augment those strategies by incorporating the ideas of Margin Sampling [7] to VU and Entropy Sampling [4] to UR respectively. This results in two new query strategies, namely *Margin Sampling with Variable Uncertainty* (MSVU) and *Entropy Sampling with Uncertainty using Randomization* (ESUR), which perform better particularly in steam mining context. These two augmented strategies are described in detail below.

3.1 Margin Sampling with Variable Uncertainty (MSVU)

The MSVU algorithm is presented below. The main difference with VU is in lines 3 and 4. In line 3, we calculate the minimum margin for two promising class labels determined by the classifier. And in line 4, this minimum margin is compared with the threshold θ.

These modifications help improve the performance of the algorithm over the original VU. The reason is that while VU considers the most possible label ignoring the information about the remaining label distribution, MSVU tackles this shortcoming by considering the difference between the two most possible labels or classes by the model.

Algorithm **MSVU** (D_t, C, β, a)
Input: (1) Incoming instance, D_t
Input: (2) Trained classifier, C
Input: (3) Budget, β
Input: (4) Adjusting step, a
Output: (1) *label* $\in$ {**true**, **false**} implies whether to ask the true label y_t
Initialization: Total labeling cost $u = 0$, initial labeling threshold, $\theta = 1.0$
1. if $(u/t < \beta)$
2. **then** budget is not exceeded,
3. $\boldsymbol{X}_C = \arg\min_{D_t} P_C(\hat{y}_1|D_t) - P_C(\hat{y}_2|D_t)$ where $\hat{y}_1$ and $\hat{y}_2$ are the first and second possible class labels respectively under model, C
4. **if**$(\boldsymbol{X}_C < \theta)$
5. **then** margin difference is below the threshold
6. u = u + 1 labeling costs increase,
7. $\theta = \theta(1 - a)$ the threshold decreases,
8. **return true**
9. **else** margin region is wider
10. $\theta = \theta(1 + a)$ make the threshold wider,
11. **return false**
12. **else** budget is exceeded
13. **return false**

3.2 Entropy Sampling with Uncertainty using Randomization (ESUR)

ESUR is a variant of entropy sampling in stream data mining context. It utilizes **sequence entropy** (SE) [4], which is defined as:

$$\Gamma^{SE}(x) = -\sum_{\hat{y}} P(\hat{y}|x:C) \log P(\hat{y}|x:C) \tag{1}$$

where $\hat{y}$ ranges over all possible label sequences for input sequence x under model C.

An algorithmic description of ESUR is presented below. ESUR is very similar to the original UR except in lines 3–5. Using Equation 1 in line 3, sequence entropy is calculated. Sequence entropy is multiplied by a random multiplier in line 4 and this resultant value is compared with threshold value in line 5. If current threshold is big enough than randomized entropy, the strategy would ask for labeling the instance. The rest of the strategy is same as UR.

These enhancements help improve the algorithm's performance over the original UR because ESUR considers the disturbance in data and also provides a more balanced coverage of the whole input space.

Algorithm **ESUR** (D_t, C, β)
Input: (1) Incoming instance, D_t
Input: (2) Trained classifier, C
Input: (3) Budget, β
Output: (1) $label \in \{\textbf{true}, \textbf{false}\}$ implies whether to ask the true label y_t
Initialization: Total labeling cost $u = 0$, initial labeling threshold, $\theta = 1$.
1. if $(u/t < \beta)$
2. **then** budget is not exceeded,
3. $\Gamma^{SE}(x)$ = Compute entropy using equation
4. $\Gamma^{SE}(x)_{randomized} = \Gamma^{SE}(x) \times \eta$,
where $\eta \in \mathcal{N}(1, \sigma^2)$ is a random multiplier
5. **if**$(\Gamma^{SE}(x)_{randomized} < \theta)$
6. **then** entropy is less than threshold
7. u = u + 1 labeling cost increases,
8. $\theta = \theta(1 - a)$ the threshold decreases,
9. **return true**
10. **else** certainty is good
11. $\theta = \theta(1 + a)$ make the uncertainty region wider
12. **return false**
13. **else** budget is exceeded
14. **return false**

4 Experimental Results

We compare the performances of five query strategies: (1) Variable Uncertainty (VU), (2) Uncertainty with randomization (UR), (3) Random, (4) Margin sampling with variable uncertainty (MSVU), and (5) Entropy sampling with variable uncertainty using randomization (ESUR). Among them, Random strategy is used as a baseline method.

For each of the five strategies, two different classifiers were used: (1) Leveraging Bagging (LB) [9] and (2) Single Classifier Drift (SCD) [10, 11]. SCD can be configured either with drift detection method (DDM) [10] or early drift detection method (EDDM) [11]. Since EDDM generally offers better outcomes, we choose to use it in our experiments.

Two sets of experiments were conducted, namely (1) Experiment on 3 prediction datasets and (2) Experiments on 3 textual datasets. All experiments are done in *Massive Online Analysis* (MOA) platform [12]. For each method, only the budget is changed for each testing instance, and default values are used for all the remaining parameters.

4.1 Experiment I: On Prediction Datasets

The three prediction datasets used are: Electricity, Forest, and Airlines [13]. These datasets are also used in the experiments of [5]. Electricity dataset is about predicting a rise or a fall in electricity demands and prices in New South Wales, Australia, provided immediate consumptions and prices in the same and neighboring regions. In Forest dataset, the task is to predict forest cover type from cartographic variables. In Airlines dataset, the task is to predict whether a given flight will be delayed or not by using supplied information of the scheduled departures.

Normal accuracy is used to evaluate the performances on the prediction datasets. The performances 5 query strategies each using 2 classifiers are summarized in Table 1.

For Airlines dataset, the performances of all strategies fluctuate almost between 65% to 50% for both classifiers. Variants of variable uncertainty strategy (VU and MSVU) outperform the variants of randomization strategy (UR and ESUR). In particular, our proposed strategy, MSVU outperforms the other strategies for both classifiers.

In the case of the Electricity dataset, MSVU outperforms other strategies for LB. With SCD as budget increases, UR shows better performance than other variable uncertainty variants. There is an accuracy fluctuation among the strategies for both of the classifiers.

All the strategies show good performance for Forest dataset. After a small budget (around 0.1), all the strategies with LB and SCD achieve just below 100% accuracy and remain stable throughout the budget change.

4.2 Experiment II: On Textual Datasets

The three textual datasets used are IMDB-E, IMDB-D, and Reuters [13]. They are also used in the experiments of [5]. IMDB (Internet Movie Database) dataset is divided into two categories. For IMDB-E (easy), only one category is considered as interesting at a time and for IMDB-D (difficult), five associated categories are interesting at a time. With the purpose of deliberately initiating concept drifts, the authors of [5] introduce three changes after 25, 50, and 75 thousand instances. In Reuters dataset, the first half of the data stream legal or judicial is considered to be relevant and in second half the share listings category was considered to be relevant. For these textual data, the labels were assigned by authors of [5].

Geometric accuracy is used to measure the performances on textual datasets. It is defined as $GA = (A_1 \times A_2 \times \ldots \times A_c)^{\frac{1}{c}}$. Here A_i is the accuracy on class i and c is the number of classes. The performances 5 query strategies each using 2 classifiers are summarized in Table 1.

The proposed strategies exhibit both high and low accuracies. For both IMDB-D and IMDB-E, ESUR attains the highest accuracy level, while VU as well as MSVU show bad accuracies. However, opposite behavior is observed in the case of Reuters dataset: MSVU achieves the highest geometric accuracy, while ESUR receives the lowest accuracy.

The variants randomization strategy present dominating performances. Among them, ESUR shows highest accuracy in all datasets. At the change in 50 thousand instances, all strategies receive their respective lowest geometric accuracies and the change in 75 thousand instances, there is a rising tendency in accuracy in the case of IMDB-E dataset while a falling tendency in accuracy is shown in the case of IMDB-D dataset at this change.

Both the classifiers, LB and SCD, show almost same behavior. The variants of variable uncertainty strategy outperform the variants of randomization strategy. In the case of LB classifier, VU shows slight better performance than MSVU. On the other hand, for SCD, MSVU shows slightly better result than VU.

Table 1. Summary of results on 3 prediction datasets (Experiment I) and 3 textual datasets (Experiment II). Highest scores are highlighted in red for LB and blue for SCD. Our proposed methods (MSVU and ESUR) provide better results than the original VU and UR methods [5] do in 9 out of 12 test cases.

	Prediction Datasets Average Accuracy (%)						Textual Datasets Average Geometric Accuracy (%)					
	Airlines		Electricity		Forest		IMDB-D		IMDB-E		Reuters	
	LB	SCD	LB	SCD	LB	SCD	LB	SCD	LB	SCD	LB	SCD
MSVU	63.77	64.52	76.45	80.62	95.26	96.36	36.61	33.70	46.96	47.80	93.27	64.56
ESUR	59.24	58.3	66.29	73.50	94.55	96.23	45.87	45.71	50.37	52.10	82.99	44.98
Random	61.58	60.11	73.18	79.63	94.67	96.10	43.51	41.56	49.25	51.37	88.70	54.47
VU	63.20	63.74	75.81	80.42	95.24	96.41	36.74	34.41	46.85	48.00	93.21	65.77
UR	61.70	61.89	74.87	78.87	95.35	96.38	42.29	41.14	48.75	50.08	89.24	56.26

5 Conclusion and Future Works

The results of the experiments described here show that ESUR perform well with the remote changes and IMDB datasets. The same is also true for MSVU for close changes as well as Airlines, Electricity, and Reuters datasets. In comparison with VU and UR [5], the proposed augmented MSVU and EUSR strategies outperform them in the majority of cases. Future research directions include designing new strategies which have the ability to tackle both close and remote changes. A more comprehensive study could also be conducted in which a larger number of base classifiers are deployed.

References

1. Bin, S., Yuan, L., Xiaoyi, W.: Research on data mining models for the internet of things. In: Proc. 2010 International Conference on Image Analysis and Signal Processing (IASP). (2010) 127–132
2. Tripathy, A.K., Adinarayana, J., Merchant, S.N., Desai, U.B., Ninomiya, S., Hirafuji, M., Kiura, T.: Data mining and wireless sensor network for groundnut pest/disease precision protection. In: Proc. 2013 National Conference on Parallel Computing Technologies (PARCOMPTECH). (2013) 1–8
3. Faisal, M.A., Aung, Z., Williams, J., Sanchez, A.: Data-stream-based intrusion detection system for advanced metering infrastructure in smart grid: A feasibility study. IEEE Systems Journal (2014) in press
4. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proc. 2008 Conference on Empirical Methods on Natural Language Processing (EMNLP). (2008) 1070–1079
5. Žliobaitė, I., Bifet, A., Pfahringer, B., Holmes, G.: Active learning with evolving streaming data. In: Proc. 2011 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) Part III. (2011) 597–612
6. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proc. 17th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). (1994) 3–12
7. Scheffer, T., Decomain, C., Wrobel, S.: Active Hidden Markov Models for information extraction. In: Proc. 4th International Conference on Advances in Intelligent Data Analysis (IDA). (2001) 309–318
8. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379–423
9. Bifet, A., Holmes, G., Pfahringer, B.: Leveraging bagging for evolving data streams. In: Proc. 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) Part I. (2010) 135–150
10. Gama, J., Medas, P., Castillo, G., Rodrigues, P.: Learning with drift detection. In: Proc. 17th Brazilian Symposium on Artificial Intelligence (SBIA). (2004) 286–295
11. Baena-García, M., Campo-Avila, J.d., Fidalgo, R., Bifet, A., Gavaldà, R., Morales-Bueno, R.: Early drift detection method. In: Proc. 4th International Workshop on Knowledge Discovery from Data Streams (IWKDDS). (2006) 77–86
12. Bifet, A., et al.: Massive Online Analysis, release: 2012.03. `http://moa.cs.waikato.ac.nz` (2012)
13. Bifet, A., Kirkby, R.: MOA (Massive Online Analysis) datastream. `http://sourceforge.net/projects/moa-datastream/files/Datasets/Classification/` (2012)