Survey on Active learning

Advanced Machine Learning Course

DIBRIS, University of Genova
Akshi Sharma

Abstract

In many real-world scenarios, we may have data in abundance but it is mostly unlabelled. Unless the data is labelled, it in essence does little good for all ML applications that rely on a supervised learning approach. To label this abundantly available unlabelled data is exhausting, time consuming and expensive with not proportional benefits. Active learning algorithms aim at minimizing the amount of labelled data required to achieve the goal of the machine learning task at hand by strategically selecting just the more relevant data instances to be labelled by the expert. By doing this, the active learner hopes to get high accuracy using as few labelled instances as possible, all the while minimizing the cost of labelling data. It exhibits satisfactory classification performance with a subset of samples than those required by conventional passive learning methods. This paper gives a brief survey of active learning; its need, working, selection methods, query strategies etc.

Introduction

With the boom of IoT devices data has only gotten cheaper to collect and store, so much so that data scientists are left with unconquerable amount of data to deal with. And it is increasing by the second leaving data professionals fending for themselves in an ocean of raw data. Even though advances in machine learning require large volumes of data and easy access to large amounts of data should be good news. But it has raised a brand new challenge: the ability to obtain sufficient labelled data for modelling purposes.

The new bottleneck in machine learning nowadays is not about the collection of data anymore, but about the speed and accuracy of the labelling process. The performance of a model majorly depends on the quality of data used to train it. Labelled data can be expensive to obtain, and frequently requires laborious human effort. Active learning can be defined as the process of prioritising the data that needs to be labelled in order to have the highest impact to training a supervised model.

What is Active Learning?

In traditional supervised learning system, labelled instances are accumulated by gathering large amount of data, which is then randomly sampled and this large dataset is used to train a model. The key hypothesis in active learning is that if a learning algorithm can choose the

data it wants to learn from, it can perform better than traditional methods with substantially less data for training [4]. The fundamental belief behind this hypothesis is that not all data are equal; some instances are more important and crucial for the actual learning depending on what exactly the machine is trying to learn. Active learning can be used in situations where the amount of data is too large to be labelled and some priority needs to be made to label the data in a smart way. An additional advantage of active learning methods is that they can often help in the removal of noisy instances from the data, which can be beneficial from an accuracy perspective [6].

Active learning is part of the so-called Human-in-the-Loop (HITL) machine learning paradigm. The idea behind HITL is to combine human and artificial intelligence to solve various data-driven tasks. Active learning algorithm is based on the simple yet powerful assumption that not all data is equally impactful on the model and that all data is not learned at the same pace by the model. An algorithm that follows this methodology is referred to as an **active learner**.

How does Active Learning Work?

Active learning systems attempt to overcome the labelling bottleneck by asking *queries* in the form of unlabelled instances to be labelled by an *oracle* (e.g., a human annotator) [4]. The oracle (the source of the ground truth labels, e.g. the human expert) supplies the model with some labelled data. The job of the query system is to pose queries to the oracle for labels of specific records [6].

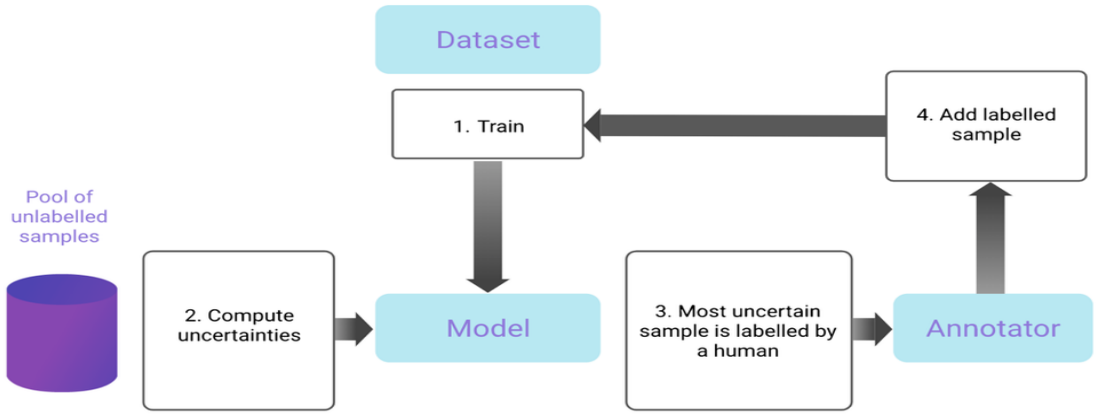

Figure 1

Scenarios

Active Learning can be implemented through different **scenarios**. Decision to query a specific label boils down to the trade-off between the cost of collecting that information and the gain from obtaining the label. In practice, that decision can be motivated by the budget or the labelling bill. Overall, three different categories of active learning can be listed:

- **Membership Query Synthesis**: In this scenario, the learner generates its own instances from the entire space for labelling. The key here is that the learner many actually construct instances from the underlying space, which may not be a part of any actual pre-existing data [6]. Query synthesis is reasonable for many problems, but labelling such arbitrary instances can be awkward if the oracle is a human annotator (in cases of speech, synthesized instances can turn out to be gibberish) [4].

Figure 2

- **Stream-Based Selective Sampling**: The key assumption is that getting an unlabelled instance is free [4]. The active learner is presented with a stream of unlabelled instances, from which the learner picks an instance for labelling by the expert [2]. As the model is being trained and is presented with a data instance, it decides if it wants to ignore the sample, or query its label.

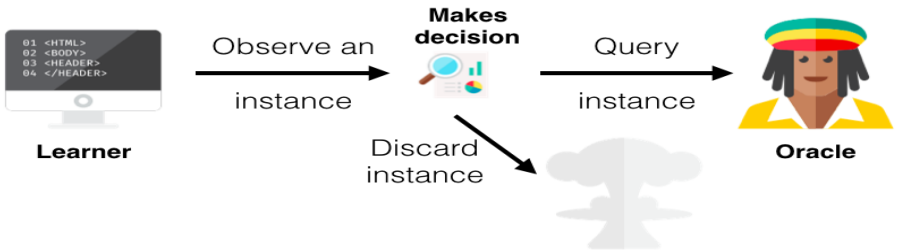

Figure 3

- **Pool-Based sampling**: Here the assumption is that there is a small set of labelled data and a large **pool** of unlabelled data [4], as with the stream-based selective sampling. The active learner is initially trained on a labelled subset of the data which generates a first version of the model which is subsequently used to identify which instances would be the most beneficial to inject in the training set for the next iteration.

Figure 4

Query Strategy

The approach used to determine which data instance to label next is referred to as a **querying strategy.** There are multiple approaches studied in the literature on how to prioritise data points when labelling and how to iterate over the approach.

- **Uncertainty Sampling-** Here the model wants to get the labels for the samples whose label assignment it is most uncertain about. This can take a few different forms:

  - ***Least confidence***: difference between the most confident prediction and 100% confidence. Samples for which the model has the lowest confidence in their most likely class are chosen. It takes the highest probability for each data point's prediction, and sorts them from smaller to larger. The actual expression to prioritise using least confidence would be:

$$s_{LC} = \underset{x}{\operatorname{argmax}} \left(1 - P(\hat{y}|x)\right)$$

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x)$$

  - ***Margin sampling***: to overcome the drawback of LC – considering only the most probable label and disregarding other label probabilities - Margin sampling takes into account the smallest difference between the top and the second most probable label of the sample. The algorithm selects the instances where the margin between the two most likely labels is narrow, i.e. they lie in between two classes in feature space, so annotating them helps in locating the decision boundary for the classifier.

$$s_{MS} = \underset{x}{\operatorname{argmin}} \left(P(\hat{y}_{max}|x) - P(\hat{y}_{max-1}|x)\right)$$

  - ***Entropy sampling***: In order to utilize all the possible label probabilities, entropy is used as it is a measure of randomness. Entropy in an information-theoretic measure that represents the amount of information needed to "encode" a distribution [4]. The entropy formula is applied to each instance and the instance with the largest value is queried.

$$s_E = \underset{x}{\operatorname{argmax}} \left(- \sum_i P(\hat{y}_i|x) \log P(\hat{y}_i|x) \right)$$

In addition to uncertainty sampling, there are other possible approaches, such as:

- ***Query-by-committee -*** It involves training multiple models, and selecting samples that these classifiers disagree about the most. At an intuitive level, the query-by-committee method achieves similar heterogeneity goals as the uncertainty sampling method, except that it does so by measuring the differences in the predictions of different classifiers, rather than the uncertainty of labelling a particular instance [6]. The core idea behind this framework is minimizing the version space, which is the set of hypotheses that are consistent with the current labelled training data [4].There is no general agreement on the appropriate committee size to use, which may vary by model class or application.

- ***Expected model change:*** This strategy aims at identifying the instances that would lead to the greatest change in the current model if we were to find out its label. Specifically, the instance that results in the greatest change in gradient of the objective function with respect to the model parameters is used [6]. The basic idea is that it prefers instances that are likely to most influence the model/parameters, irrespective of the label of the resulting query. This approach has been shown to work well in empirical studies, but can be computationally expensive if both the feature space and set of labels are very large [4].

- ***Expected error reduction:*** Instead of looking for the greatest gradients, the change in the validation error is computed. Query the instances that would most reduce error. The point is to measure how much its generalization error is likely to be reduced. The aim is to estimate the expected future error that would result if some new instance x is labelled and added to trained data, and then select the instance that minimizes that expectation [4].Because minimizing an expected loss function usually doesn't have a close form solution, a variance reduction approach is used as a proxy.

Challenges

More than 9 researchers out of 10 who have attempted some work involving Active Learning claim that their expectations were met either fully or partially [8]. That's very encouraging, but it also implies that there are cases where it does not work at all. The least confidence method (although not the only one) tends to gravitate towards the outliers. One solution to this is could be to iterate between different query strategies, e.g. from least confidence to random, then to margin sampling, etc. Active Learning is a category of algorithms which can be vulnerable to biases, because of the initial prejudice they might make by the patterns identified when the model trained on a small dataset.

But the most significant issue with Active Learning is that it still not well understood or researched on. There is very little work, for example, on predicting ahead of time if a specific

task or dataset is particularly prone to benefit from an active learning approach. Not to mention that there is no theoretical framework for Active Learning hyper parameter tuning, right choice of querying strategy or stopping criteria.

Conclusions

Active learning enables the application of machine learning methods to problems where it is difficult or expensive to acquire expert labels. Active learning can be the answer to the woes of all machine learning problems which suffer from the constraints of labelled data. One of the most popular areas where active learning can be a boon is natural language processing (NLP) as it requires lots of labelled data (for example, Part-of-Speech Tagging, Named Entity Recognition) and the cost to label this data is very high.

With efficient labelling becoming an ever-more critical component of Machine Learning, it is safe to assume that a lot more research will be done on the topic in the coming years .Active learning may not be a one-size-fits-all, but it is a heavily under-utilized technique.

References

[1] L. Sun and X. Wang, "A survey on active learning strategy," 2010 International Conference on Machine Learning and Cybernetics, Qingdao, 2010, pp. 161-166, doi: 10.1109/ICMLC.2010.5581075.

[2] Krishnakumar, Anita. (2007). Active Learning Literature Survey.

[3] Settles, Burr & Editor, I & Guyon, G & Cawley, Gavin & Dror, V & Lemaire, A & Statnikov,. (2011). From Theories to Queries: Active Learning in Practice. 16.

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

[5] https://www.datacamp.com/community/tutorials/active-learning

[6] Aggarwal, C. (2014). Chapter 22 Active Learning : A Survey.

[7] https://blog.scaleway.com/active-learning-some-datapoints-are-more-equal-than-others/

[8] https://www.kdnuggets.com/2018/10/introduction-active-learning.html

[9] https://towardsdatascience.com/active-learning-in-machine-learning-525e61be16e5

[10] https://towardsdatascience.com/are-you-spending-too-much-money-labeling-data-70a712123df1