MJRET

OPEN ACCESS

Multidisciplinary Journal of Research in Engineering and Technology

www.mjret.in

eISSN:2348-6953

ACTIVE LEARNING METHODS FOR LABELLING DATASETS

Atharv Wani, Gorla Charan Sai Chowdary, Merin Meleet, Dr. Anala M R

Department of Information Science and Engineering R V College of Engineering

Mysore Road, Bengaluru

Abstract: *Data scientists are challenged with more data than they will ever be able to analyse as data collection and storage costs continue to drop. The most fascinating developments in machine learning require vast amounts of data. But it also creates a new challenge for the machine learning community, as all supervised learning-based machine learning applications remain practically useless without labelled data. Labelling large datasets has become a vital challenge. A specific instance of Supervised Machine Learning is Active Learning. By actively choosing the important data points, this method builds a high-performance classifier thus minimizing the size of the training dataset.*

Keywords: *Active learning, Supervised Learning, Oracle, Least confidence, Margin sampling, Entropy.*

1. INTRODUCTION

Active learning is part of machine learning in which a learning algorithm can query a user interactively to label data with the outputs which are desired. In active learning, the algorithm intelligently selects the subset of examples to be labelled next from the pool of unlabeled data. The basis behind the active learning algorithm concept is that an ML algorithm could possibly reach a higher level of accuracy while using a smaller number of training labels if it was allowed to choose the data points from which it wants to learn from. Identifying the best instances for a model to learn can happen at two different times: either before the model is even being built referred to as prioritisation, or while the model is trained which is called Active Learning. Therefore, the algorithm is allowed to interactively pose queries during the stage of training. These queries are usually in the form of unlabeled data points which request a human annotator to label the data point. This makes active learning part of the human-in-the-loop archetype, where it is one of the most powerful examples of success. Active Learning is really very well-motivated in many contemporary ML issues, especially

when labels are challenging, time-consuming, or expensive to gather, despite the fact that it receives little coverage in well-known ML blogs. In active learning, the learning algorithm is given the freedom to actively choose from a pool of previously unlabeled examples the subset of examples that will be labelled next. The underlying idea behind the concept is that, if given the freedom to select the data it wants to learn from, a Machine Learning algorithm might be able to improve accuracy while requiring less training labels. An algorithm like this is known as an active learner. During the training process, active learners are permitted to dynamically pose queries, typically in the form of unlabeled data instances. Active learners are allowed to pose queries during the training process, usually in the form of unlabeled data instances to be labelled by what is called an oracle, usually a human annotator.

By actively choosing the important data points, this method builds a high-performance classifier while keeping the minimised size of the training dataset. Data science teams can save a lot of time and compute by selecting which data points to classify first when labelling. The model needs to be trained using a limited amount of labelled data which was actually labelled by an oracle. The model won't be perfect, of course, but it will provide us some insight into which regions of the parameter space should be tagged first to make it better. The model is used to predict the class of each subsequent unlabeled data item after it has been trained. Each unlabeled data point is assigned a score depending on the model's prediction. Least confidence, Margin Sampling, and entropy computations are three methods for determining a priority score for each data point.

1.1 Least confidence

This approach is probably the simplest. After calculating the probabilities of each data point and sorting them from smallest to largest. The data points with the highest probability are included in the training set for the next iteration.

$$S_{LC} = \operatorname{argmax}(1-P(\hat{Y}|X))$$

$$\hat{Y} = \operatorname*{argmax}_{Y}(1-P(Y|X))$$

1.2 Margin Sampling

In this approach after calculating the probabilities of each data point the difference between the highest probability and the second highest probability is taken into consideration.

$$SMS = \operatorname*{argmin}_{X}(P(\hat{Y}max|X) - P(\hat{Y}max\text{-}1|X))$$

1.3 Entropy

As we know Entropy is a measure of disorder in a system. If there is more disorder, then higher is the entropy, and if disorder is low, lower is the entropy. This idea can be applied to measure the certainty of the model. In the case of high entropy, this would imply that the model equally distributes the probability for all classes because it is completely uncertain to which class that data point belongs. It is therefore obvious to prioritise data points with higher entropy to the ones with lower entropy.

2. HYPOTHESIS

The core hypothesis of active learning is that, with far less training data than conventional approaches, a learning algorithm can outperform them if it is given the freedom to select the data it wants to learn from.

The technique of prioritising the data that has to be labelled in order to have the most influence on training a supervised model is known as active learning. When there is too much data to label and smart labelling needs to be prioritised because there is too much data to label, active learning can be employed.

3. LITERATURE REVIEW

William H, Tim Genewein , Andreas [5] worked on The power of ensembles for active learning in image classification where They found ensembles perform better and lead to more calibrated predictive uncertainties, which are the basis for many active learning algorithms. Additionally, they showed results on a large, highly class-imbalanced diabetic retinopathy dataset.

Ozan Sener, Silvio Savarese [7] worked on Active learning For Convolutional neural networks: A Core-Set approach. The objectives carried out in this work include In order to tailor an active learning method for the batch sampling case, they decided to define the active learning as a core-set selection problem. Core-set selection problem aims to find a small subset given a large labelled dataset such that a model learned over the small subset is competitive over the whole dataset.

Sanjoy Dasgupta of University of California [8] worked on Analysis of a greedy active learning strategy. The major gaps identified include a greedy approach is not optimal because it doesn't take into account the way in which a query reshapes the search space – specifically, the effect of a query on the quality of other queries.

Yarin Gal, Riashat Islam, Zoubin Ghahramani in [9] worked on Deep Bayesian Active Learning with Image Data. The results obtained in this paper include system is able to achieve 5% test error on MNIST with only 295 labelled images without relying on unlabelled data (in comparison, 835 labelled images are needed to achieve 5% test error using random sampling – requiring an expert to label more than twice as many images to achieve the same accuracy), and achieves 1.64% test error with 1000 labelled images.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin in [14] worked on Cost-Effective Active Learning for Deep Image Classification.The objectives carried out in this work include to apply our CEAL framework to deep image classification tasks by progressively selecting complementary samples for model updating. The gaps identified in this research include the Need framework on more challenging large-scale object recognition tasks (e.g., 1000 categories in ImageNet).

Melba M. Crawford, Fellow IEEE, Devis Tuia, Member IEEE, and Hsiuhan Lexie Yang in [13] worked on Active Learning: Any Value for Classification of Remotely Sensed Data.The objectives carried out in this work include spatially contiguous pixels can be useful in contextual classifiers, they are typically overrepresented, and only a subset of these pixels contributes effectively to the development and performance of the classifier. The gaps identified in this research include although the assumption that pixel labels can always be obtained is not always practical or even possible in remote sensing applications.

4. METHODOLOGY

The methodology contains three major steps which are described as follows:

4.1 Gathering Data and Annotations

Here we have a MNIST Handwritten digit dataset of 42000 images. Which had different images distributed in 10 classes.

Size of overall Dataset - 42000 Images Size of a Image - 28X28 grey-scale image Shape of a Data-point - (28, 28, 3)

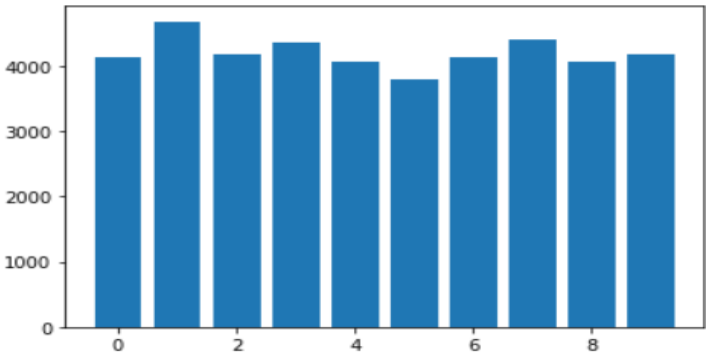


Fig 4.1 Total no. of instances in each class

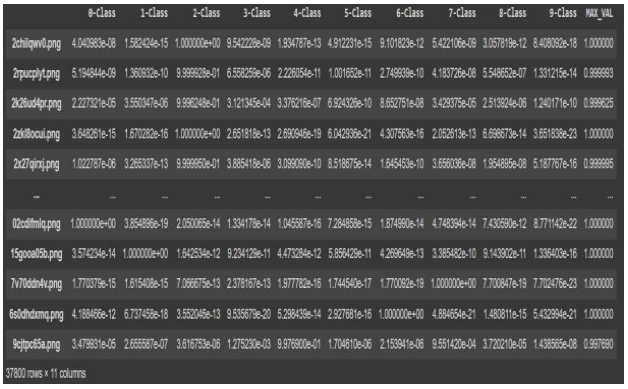
There are a total of 10 classes [0-9] and the count of images in each class are plotted.
{ 0.0: 4132, 1.0: 4684, 2.0: 4177, 3.0: 4351, 4.0: 4072, 5.0: 3795, 6.0: 4137, 7.0: 4401 ,8.0: 4063, 9.0: 4188 } Adding 10% images of each class in the training set for initial iteration and keeping rest in the testing set. This is also called the initial annotated set.
Overall size of Training-set 4200 (10% of 42000) Overall size of Testing-set 37800 (rest 90%)

4.2 Model building and training

Convolutional Neural Network, or CNN for short, is a model that has gained a lot of attention recently due to its utility. Multilayer perceptrons are used by CNN to perform computational tasks. Comparatively speaking to other image classification methods, CNN employs a little amount of pre-processing. This indicates that filters that were manually created for traditional algorithms are now used by the network to learn. CNNs are therefore the ideal choice for tasks involving image processing. The below CNN model gives 85.28% Accuracy on initial iteration and we now with SOFTMAX function will get prediction scores for each class. Here we employ the least confidence method to calculate priority scores for each data point.

A vector of numbers is transformed into a vector of probabilities via the mathematical operation known as Softmax, where the probability of each value are directly proportional to the relative scale of each value in the vector.

This way we have a prediction value given by the model to each class for every data point.



	0-Class	1-Class	2-Class	3-Class	4-Class	5-Class	6-Class	7-Class	8-Class	9-Class	MAX_VAL
2chiiqwv0.png	4.040883e-08	1.582424e-15	1.000000e+00	9.542228e-09	1.934787e-13	4.912231e-15	9.101823e-12	5.422106e-09	3.057819e-12	8.408092e-18	1.000000
2rpucpiyt.png	5.194844e-09	1.360932e-10	9.999928e-01	6.558259e-06	2.226054e-11	1.001652e-11	2.749939e-10	4.183726e-08	5.546652e-07	1.331215e-14	0.999993
2k26ud4pr.png	2.227321e-05	3.593047e-06	9.996248e-01	3.121345e-04	3.376216e-07	6.924326e-10	8.662751e-08	3.429375e-05	2.513924e-06	1.240171e-10	0.999625
2zki8ocui.png	3.648261e-15	1.670282e-16	1.000000e+00	2.651818e-13	2.690946e-19	6.042936e-21	4.307563e-16	2.052613e-13	6.698673e-14	3.651838e-23	1.000000
2x27qinxj.png	1.022787e-06	3.265337e-13	9.999650e-01	3.885418e-06	3.099090e-10	8.518675e-14	1.645453e-10	3.656036e-08	1.954885e-08	5.187767e-16	0.999965
...	...	...	...	...	...	...	...	...	...	...	...
02cdlfmiq.png	1.000000e+00	3.854896e-19	2.050065e-14	1.334178e-14	1.045687e-16	7.284858e-15	1.874986e-14	4.748394e-14	7.430590e-12	8.771142e-22	1.000000
15goaa05b.png	3.574234e-14	1.000000e+00	1.642534e-12	9.234129e-11	4.473284e-12	5.856429e-11	4.269649e-13	3.385492e-10	9.143902e-11	1.336403e-16	1.000000
7v70ddn4v.png	1.770379e-15	1.615408e-15	7.086675e-13	2.378167e-13	1.977782e-16	1.744540e-17	1.770362e-19	1.000000e+00	7.700847e-19	7.702476e-23	1.000000
6s0dhdxmq.png	4.188466e-12	6.737459e-18	3.552046e-13	9.535679e-20	5.298439e-14	2.927681e-16	1.000000e+00	4.884654e-21	1.480811e-15	5.432994e-21	1.000000
9cjtpc65a.png	3.479931e-05	2.655587e-07	3.616753e-06	1.275230e-03	9.976900e-01	1.704610e-06	2.153941e-06	9.551420e-04	3.720210e-05	1.438565e-08	0.997690

37800 rows × 11 columns

Fig 4.2 Priority scores for Testing set

4.3 Performing Active learning iterations

Now after getting prediction scores we perform the first active learning iterations using LEAST CONFIDENCE and take out the data points where the model is 100% sure about its prediction. This way human/oracle inclusion is not needed and the model itself comes to know about the data points that are to be added in the training set to create the most competitive/smart training set for prediction on the testing set in subsequent iteration.

```
print(newtestingset.shape)
print(newtrainingset.shape)
```

```
(30238, 28, 28, 3)
(11762, 28, 28, 3)
```

```
print(testingset.shape)
print(trainingset.shape)
```

```
(37800, 28, 28, 3)
(4200, 28, 28, 3)
```

Fig 4.3 1st iteration of Active learning

The newtestingset and newtrainingset is created after the first iteration whose shape is compared in the above Fig 4.3 with the trainingset and testingset prior to active learning iteration. Here we can see that after each iteration there are certain data points which will get into the training-set from the testing-set. After performing the 1st iteration accuracy In the subsequent iterations we can see that testing-set size decreases and the training-set size increases. The required max-value for the data point to be included in the training-set is ‘1’ across the vector generated by the model for the data point.

5. EXPERIMENTAL RESULTS

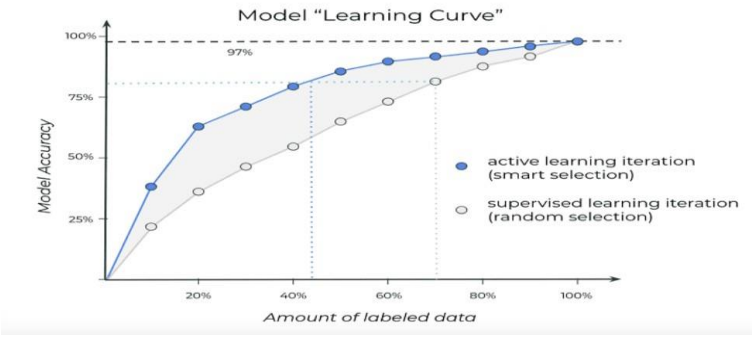


Fig 5.1 Comparison of Active learning and supervised learning iterations [3]

The accuracy of the active learning model is checked using the true labels which are taken from the MNIST Handwritten Digit Dataset's train labels. Our aim with this experiment was to check the real life implementation and check the feasibility of an active learning model. After the first iteration it was seen that accuracy was 90% and in subsequent 3-4 iterations it went to 98% thus the results show that the accuracy in further active learning iterations will definitely converge to 100% as more and more points are labelled.

6. CONCLUSION AND FUTURE WORK

The reality that data is becoming cheaper or less expensive to gather but more challenging or expensive to label for training has undoubtedly contributed to the growth of the field of active learning in machine learning. There has been a lot of work done during the last two decades. This has produced a lot more evidence that various applications can successfully minimize the number of labelled instances required to train accurate models. Using these as a basis, the current research surge appears to be focused on using active learning techniques in everyday life, which has given rise to a number of significant problem variants and practical difficulties. These problems touch on a variety of interdisciplinary subjects, including statistics, cognitive science, and human-computer interaction. There can be a lot of other techniques which can be used to identify exact data points which can be included in the training set so as to get better results on the training set.

REFERENCES

[1] C. Persello and L. Bruzzone, "Active and Semisupervised Learning for the Classification of Remote Sensing Images", IEEE Trans Geoscience and Remote Sensing, vol. 52, no. 11, pp. 6937-6956, Nov. 2019.

[2] Z. Wang, B. Du, L. Zhang, L. Zhang and X. Jia, "A Novel Semisupervised Active-Learning Algorithm for Hyperspectral Image Classification", IEEE Transactions on Geoscience and Remote Sensing, 2019.

[3] Jennifer Prendki, VP of Machine Learning, Figure Eight "Introduction to Active Learning" SunJackson Blog.

[4] K. Wang, D. Zhang, Y. Li, R. Zhang and L. Lin, "Cost-Effective Active Learning for Deep Image Classification", IEEE Trans. Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2591- 2600, Dec. 2018

[5] William H., Tim Genewein , Andreas "The power of ensembles for active learning in image classification", IEEE/CVF Conference on Computer Vision and PatternRecognition, 2018.

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

[7] Ozan Sener, Silvio Savarese "Active Learning for Convolutional Neural Networks: A Core-Set Approach " , ICLR Conference Blind Submission , 2018.

[8] Sanjoy Dasgupta " Analysis of a greedy active learning strategy" , Advances in Neural Information Processing Systems 17 (NIPS 2004)

[9] Yarin Gal, Riashat Islam, Zoubin Ghahramani "Deep Bayesian Active Learning with Image Data" Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1183-1192, 2017.

[10] Z. Wang, B. Du, L. Zhang, L. Zhang and X. Jia, "A Novel Semi supervised Active-Learning Algorithm for Hyperspectral Image Classification", IEEE Trans. Geoscience and Remote Sensing, vol. 55, no. 6, pp. 3071-3083, June. 2019

[11] Z. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers", IEEE Trans. Knowledge and Data Engineering, vol. 17, no. 11, pp. 1529-1541, Nov. 2020

[12] S. Ravi and H. Larochelle, "Meta-learning for batch mode active learning", International Conference on Learning Representations workshop, 2018.

[13] Melba M. Crawford; Devis Tuia; Hsiuhan Lexie Yang "Active Learning: Any Value for Classification of Remotely Sensed Data" , Proceedings of the IEEE (Volume: 101, Issue: 3, March 2013).

[14] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, Liang Lin " Cost-Effective Active Learning for Deep Image Classification", Accepted by IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 2016

[15] M. Li, R. Wang and K. Tang, "Combining SemiSupervised and active learning for hyperspectral image classification", Computational Intelligence and Data Mining (CIDM) 2013 IEEE Symposium on, pp. 89-94, 2019.