Active Learning from Data (Supervised Learning)

Harshada Raut, MSc in Data Science, University of Essex, U.K

Abstract—The main aim of the paper is to test a new approach to deal with large data sets using Active learning (Supervised learning).In real life incidence, 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. In this paper, we will disscuss the goal of this advance learning approach is to speed along the learning process, especially when we have a large labeled dataset to practice traditional supervised learning methods. In this paper, an approach of how active leaning will be used for two different dataset which were collected from different domains will be discussed along with data description, experiments, impact of data sizes, limitations, challenges and their results will be presented. There results of the active learning algorithms will also be compared with cross validation and its performances will be discussed.

Index Terms—Active learning, classification, supervised learning, cross validation, rank-batch , stream-pooling.

1 Introduction

We are living in 21st century where it is no secret that training modern machine learning (ML) models requires large quantities of training data. To train supervised machine learning algorithms, we need:
1.Data and annotations for the data.
2.The ability to "learn" from the data, usually by optimizing a model so it fits the data and its annotations.
Most focus of the machine learning is onto creating better algorithms for learning from data. But getting useful annotated dataset is difficult task. It can be expensive, time consuming, and you still end up with problems like annotations missing from some categories.Moreover, dimensionality curse is a common engineering problem. Ineffective and unreasonable reduction of the dimensionality of the variable jeopardizes the efficiency of machine learning, the accuracy of pattern recognition and the efficiency of data mining while increasing the workload of measured data experiments to some extent. However, there are situations in which unlabeled data is abundant but manual labeling is expensive. In such a scenario, learning algorithms can actively query the user for labels. This type of iterative supervised learning is called "Active learning".It is basically used to label the unlabeled data in a smart way. In this technique, the number of examples to learn a concept are better understood when the data is much lower than the number required in normal supervised learning. With this approach, there is a risk that the algorithm is significantly overwhelmed by uninformative examples. In this paper we have dedicated to use multi-label active learning for the same.(We will use of ModAL libraries) In this paper, two different dataset from different domains have been selected. Both domains are different as one of them set is the most popular Iris flower dataset where as the wheat seeds dataset is widely used for machine classification problem. The Iris data set consists of 50 samples with one target variable which contains the three species of Iris (Iris setosa, Iris virginica and Iris versicolor).On the other hand,The wheat dataset is used to examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian containing 75 elements each, randomly selected for the experiment. While experimenting any species, Active learning techniques of Machine learning have been popularly used in intelligent systems

in recent decades. A precise classification is also used to differentiate the for getting most appropriate results.

2 Background

Different machine learning algorithms have been developed recently to tackle this problem, which mostly have been based on sample techniques, cost sensitive learning and ensemble methods.We can handle the large datasets adequately by using Logistic Regression algorithm, Cross Validation and accuracy metric from Scikit-learn. Whereas many machine learning algorithms are designed to maximize overall accuracy by default. In most cases, we have to face classification problem when we are provided with a big, unlabelled data sets and are asked to train well-performing models with them. Generally, the provided data can be to large to manually label it, and it becomes quite challenging for data teams to train good supervised models with that data. Active learning is the name used for the process of prioritising the data which needs to be labelled in order to have the highest impact to training a supervised model.[1]

In Active Learning, the algorithm itself is allowed to proactively select the subset of available examples to be labeled next from a pool regardless it is unlabeled. The fundamental concept of AL is that a Machine Learning algorithm could significantly achieve a better accuracy. To achieve this goal it will work with fewer training labels if it were allowed to choose the data it wants to learn from. Active learners are allowed to dynamically perform queries during the training process, usually in the form of unlabeled data instances to be labeled by what is called an oracle. In this we can simply initialize classifier to a random state with unlabeled data set. It is carried out under different scenarios which will be useful to label the unlabeled dataset beneficially. Firstly, The ”Stream-based selective sampling” model is being trained and is presented with a data instance.It immediately decides if the model wants/needs to see the label. The model being so expensive is a noticeable drawback. Secondly, the ”Pool-based sampling” is also the best-known one among the all in AL Algorithms. It is highly recommended as before selecting the best query or a set of best queries it attempts to evaluate the entire dataset. The active learner is usually initially trained on a fully labeled fraction of the data which generates a first version of the model.Later, this model is subsequently used to identify which instances would be the most beneficial to inject in the training set for the next iteration (or active learning loop). It comes from its memory-greediness which is considered as one of its major drawbacks. However, we have successfully used the pool based sampling on the datasets. Last but not the least,the membership query synthesis scenario might not be applicable to all cases as it implies the generation of synthetic data. In this scenario, the learner is allowed to construct its own examples for labeling.[2] These are the three main technique in active learning that can be used to optimise the data points chosen for labelling and training a model based on them. Therefore, we will discuss how AL in machine learning is a smart algorithm and widely used for unlabeled data in prioritise when labelling to save data science teams large amounts of time of computation with more accuracy.

3 Methodology

Two data sets were downloaded from the Kaggle web-site with one target variable Iris dataset and Wheat datasset.These datasets were taken from the JOURNAL OF DATA SCIENCE, APRIL 2020 and were in comma separated value (.csv) format and relatively small in size. Iris data has 150 samples/instances whereas the wheatdataset has 199 samples. The data was first checked for missing values and there were none found. Algorithms were developed with Google Colab using the python programming language and scikit-learn software libraries that contain many routines that are suitable for machine larning.

3.1 Iris Dataset

We will begin the analysis by exploring the dataset. Firstly, we looked for the missing

values but as there were none we proceded with the findings. The data set consists of 50 samples from each of three species of Iris(Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The use of this data set in cluster analysis and classification algorithms. since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information.There are Six attributes in the dataset with one target variable SPECIES as follows: Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Species

Furthermore, we begin the analysis by counting the sample frequency as shown in the figure1. and we will also consider the relation between the variables from heat map as shown in fig2.It is clear that petal length is highly co related to that of sepal length and vice versa. Therefore, we concluded that the longer the sepal length of a flower the petal length will be more. The scales of each column in the after counting the sample we can clearly see that there is no difference. Hence we will not perform normalization before working on the dataset.

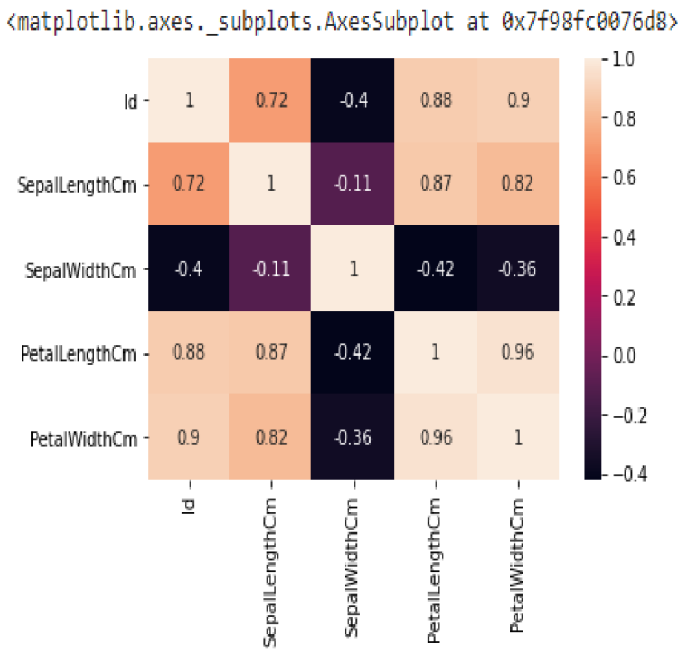

Fig. 1: Correlations of variables(HEATMAP)

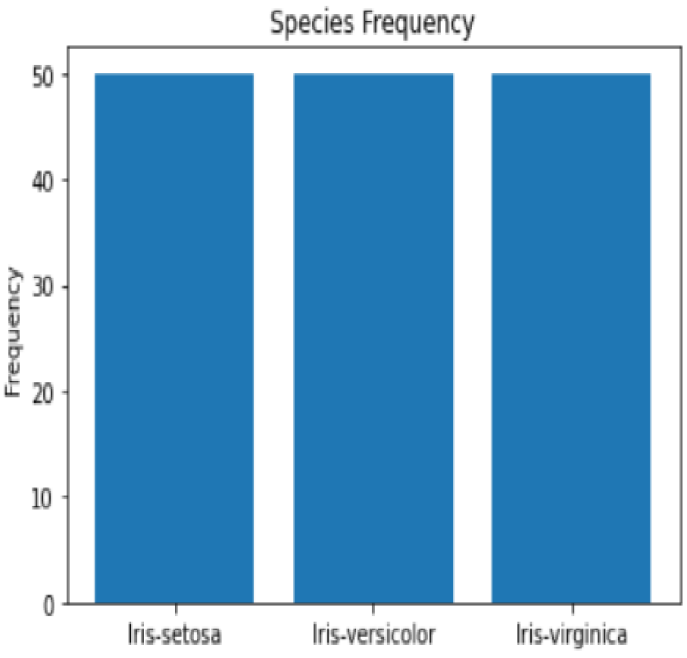

Fig. 2: Iris Sample Frequency

3.2 Wheat Dataset

The examined group of wheat Seed dataset is comprised kernels belonging to three

different varieties of wheat: Kama, Rosa and Canadian.The target variable of the dataset is Type which refered to the variety of wheat which belongs to either 1,2 or 3. It has 150 elements randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology.[3] The data set is popularly used for the tasks of classification and cluster analysis in machine learning. It has 8 attributes with one target variable as follows.
1. area A
2. perimeter P
3. compactness
4. length of kernel
5. width of kernel
6. asymmetry coefficient
7. length of kernel groove
8.Type

Moreover, we did count the sample number of each variety by using box plots as shown in the below figure there are total 68 instances of seeds belongs to kama, 66 are taken from rosa and 65 belongs to Canadian variety. However, The heatmap significantly gives the correlation between all the attributes and as shown in the figure 4 the Kernel.width of the seed has a higher impact on the Compactness of the seed. Hence, We have even plotted a graph against the kernel.width and Compactness in the analysis. After exploring the datasets we carried out the data processing.

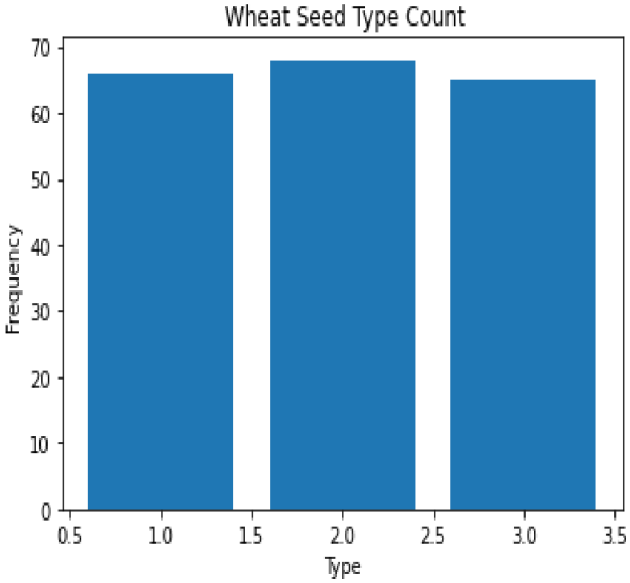

Fig. 3: Wheat Seeds Type Count

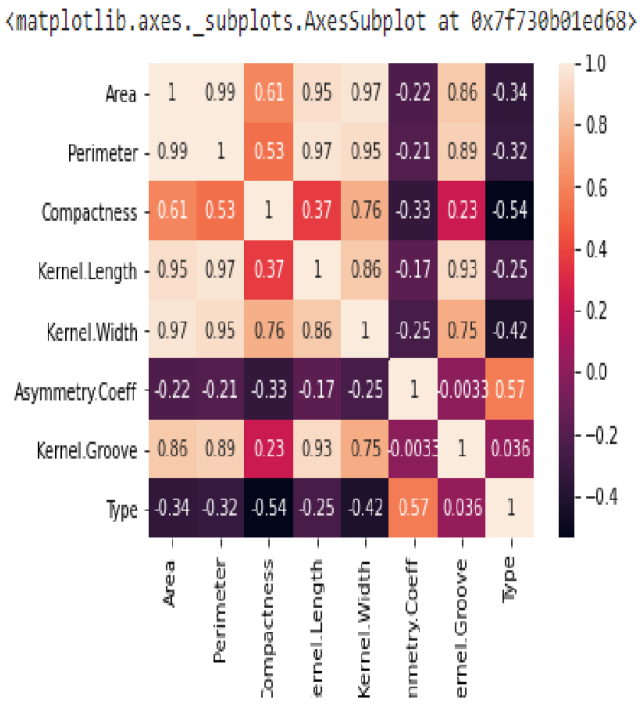

Fig. 4: Correlation between attributes using heatmap

3.3 Data Preprocessing

Data pre-processing techniques includes data preparation and data transformation into a

suitable form for data mining purposes. The goal of data pre-processing is to reduce the dimensions of the data, clean the data, find patterns in the data as we need to normalize data and remove outliers. Since we are dealing with two sets of dataset is obvious that the features have different scales and each features can be ordinal, categorical and numerical. Also, to process any data it has to be converted to numerical format for the algorithms to access. The following pre-processing methods have been applied to comparatively to all datasets during this research.

Data Cleaning: instances in the data had incomplete values, noise and outliers. The first step in data pre-processingis to correct the inconsistent data as these inconsistent data might affect the performance of our algorithm. Therefore, the data has been cleaned before passing it to algorithm Converting categorical (text) values into numerical values : Many of the variables in Census data are categorical in which are in text. The categorical variables are converted into numeric values.
Removing unrelated columns: We do not have any unrelated columns in the dataset hence we did not remove any of the variable as all are contributing to the target variable.If otherwise it is always better to remove unwanted columns.
Plot correlation matrix: The next technique which could be implemented is the correlation plot across all the variables to gain information about the correlation between each variable and the target variable also the correlation between independent variables.We used paired plots and histogram to outline the correlation between the attributes.
Label encoding: Since our dataset contains text as values in some of the rows.In our both the datasets the target value has three categories in text. Label encoder encodes the values 1,2 and 3.The number of the two classes depends on the number of classes a feature. The drawback of label encoding is that it gives weight for higher numbers and may lead to the false assumptions.

4 Experiment

To achieve the aim of the given task, each of the data sets in turn were loaded into python dataframes using the scikit-learn routine Random Forest Classifier.A random forest classifier depends upon tree-based methods for classification, and also for regression, and then a random forest classifier was fitted to the dataset.Tree based methods rely on segmenting the predictor space into regions, and then predictions can be made according to the training data classifications. We can either chose mean method to compute the average of the instances or mode can be used as well. The rules for splitting the data can be depicted in the form of a tree, and hence these approaches are known as decision tree methods.We used the The random forest procedure for 10 fold classification as it uses bagging.However, Considering the size of the datasets it wasnt feasible to obtain 10 folds hence we have set 5 folds for cross validation.But when building the decision trees only a random proper subset of the predictors is chosen from the full set The random procedure aids in decor relating the generated trees therefore we have reduces variance in the results. In order to determine the quality of random forest procedure was at predicting the class labels of unlabelled data,the k fold cross validation procedure played an important role. Furthermore, The two metrics chosen for the analysis performance were 5 different categories such as accuracy, balanced accuracy , F1 score, the mean and standard deviation of these quantities were determined over the 5 folds. Moving on to the next part of the analysis, we performed the algorithms of Active learning on both the datasets. Firstly, We classified our datasets into 2 parts training and testing equally. Then we installed modAL package to perform Rank-batched and Pool based sampling algorithm on our datasets.

4.1 Rank-batched Algorithm

We carried out the the Cardoso et al.'s ranked batch-mode sampling which not only supports batch-mode sampling (sampling methods that are built with querying multiple labels from

the unlabeled set) but also establishes a ranking among the batch in order for end-users to prioritize which records to label from the unlabeled set. One of the drawbacks of standard pool-based sampling (indeed, most sampling techniques) is their interactive nature.The rank-batch sampling methods are best fit for cases where we have an attentive human-in-the-loop for maximizing the model's performance.
This is a new approach in the active learning as the algorithm has the ability to generate an arbitrarily long query, thus making its execution less frequent. For example, a ranked query containing every available instance could be generated outside working hours. This gives us enough time as it allow to hired analysts to label instances for a full day. And alongside if desire we can update and reconstruct the query to understand it better.Hence, the algorithm is widely used not only for consuming lesser time but for improving the accuracy of the dataset as well.[4]

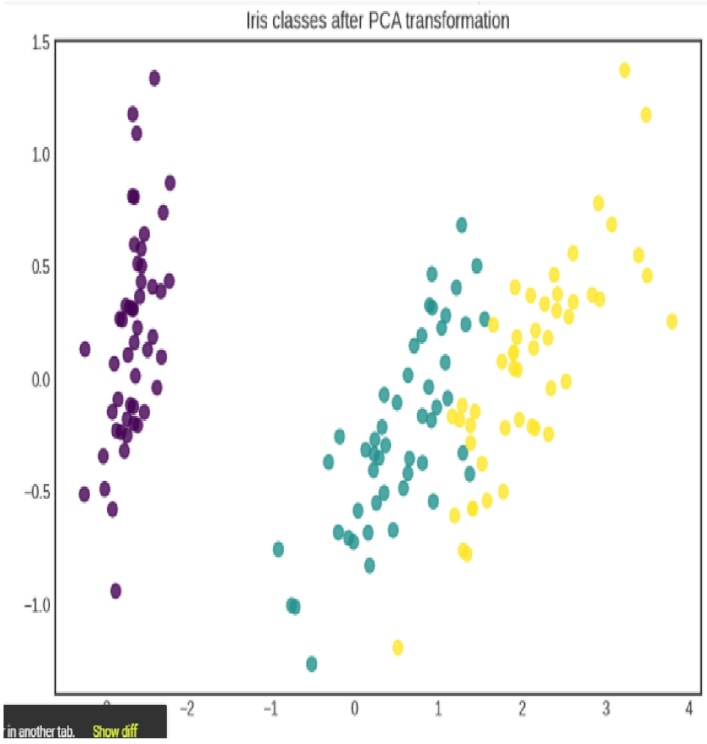

Fig. 5: PCA on iris dataset

As we can see clearly in the figure 5 we applied PCA(Principle Component Analysis) as it is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. As the ability to generalize correctly becomes exponentially harder as the dimensionality of the training dataset grows, thus the training set covers a dwindling fraction of the input space. Therefore, PCA Models also become more efficient as the reduced feature set boosts learning rates and diminishes computation costs by removing redundant features. However, As shown in the figure 6 after performing batch-rank active learning the accuracy only after 1st query has been improvised. The black dots are training points which are at distantly selected on the unlabeled dataset.

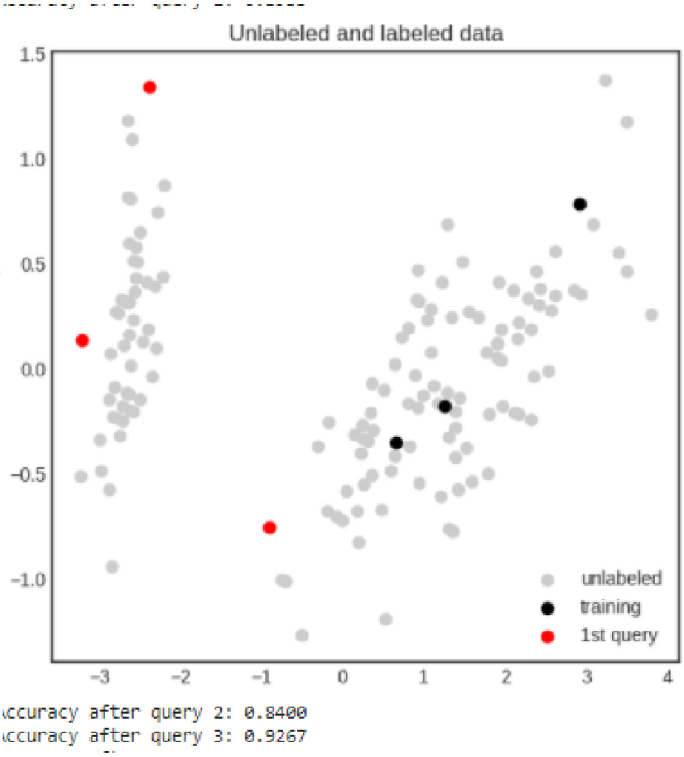

Fig. 6: Accuracy of Iris Dataset

4.2 Pool based sampling

Pool-based method is one of the hot research topics of active learning and a lot of research methods in this regard have been proposed in recent years. Most of pool-based active learning is used when the dataset can be assumed as a pool filled with unlabeled data.In this method, the most uncertain data is most informative or valuable and select it as representative to label

during learning.
We begun our analysis by splitting the dataset into our pool and test sets. These was followed by the typical 50–50 breakdown seen in most training problems. The pool is then broken out into training and validation sets. We can only select k examples to use in the training set as the rest go to the validation set. Moreover, we trained the dataset and (ignoring the validation set) then evaluate our model on the test set. Furthermore, instead of getting labels for both the validation and training sets, we only get labels for training set items. There is a most obvious reason behind it as we don't have to consult our oracle as much for labels. The other reason is we won't actually use the labels during validation. For analysis, we did the validation because it is a phase in which our algorithm attempts to predict the labels of the validation examples and outputs a value for how confident it was in its decision. As we can see in the figure 7, our model is unable to properly learn the underlying data distribution. All of its predictions are for the third class label, and as such it is only as competitive as defaulting its predictions to a single class. But that is the reason why we tune our classifier by allowing it to query 20 instances which it hasn't seen before.We will perform 20 queries on the dataset and then as shown in the figure 8 we gained more accuracy by using uncertainty sampling. Our classifier aims to reduce the amount of uncertainty in its predictions using a variety of measures with each requested query. Hence, in the end we remove that record from our pool U and record our model's accuracy on the raw dataset. However, in real life, it's very likely some valuable data don't exist in the pool, so can't guarantee to get the most valuable data from pool which may results imperfect final classifier.

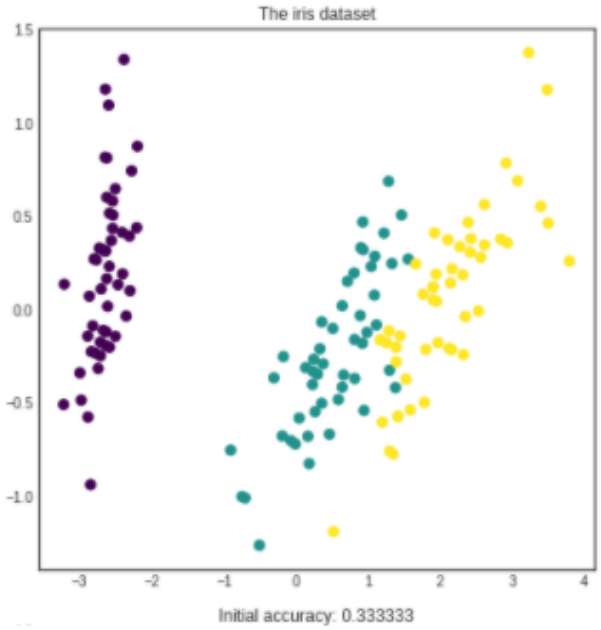

Fig. 7: Initial accuracy before pooling

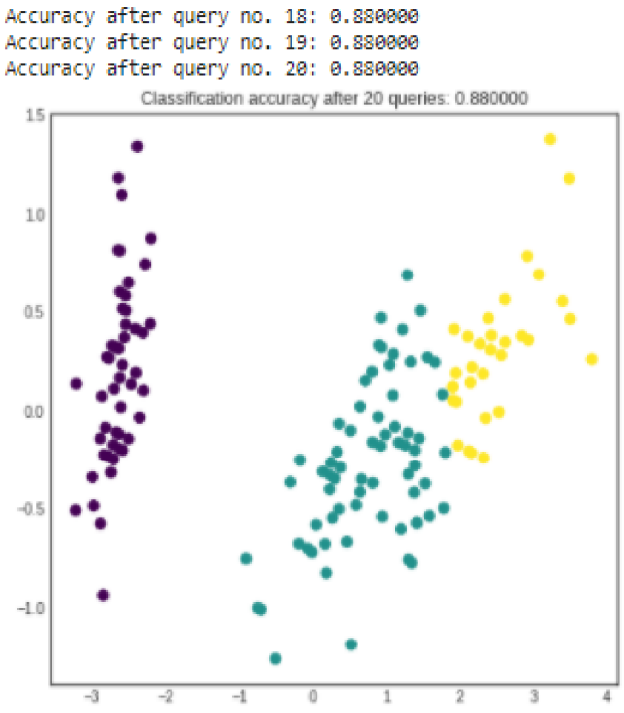

Fig. 8: Accuracy after 20 queries in pooling

5 Result

The performance metrics produced for both the data sets using the random forest algorithm

only without 5 folds stratified cross validation for Iris dataset and 10 folds stratified cross validation for Seeds dataset are shown in Table 1.

TABLE 1: 10/5 - fold Cross validations

Data set	precision	recall	F1-Score
Iris dataset	0.15	0.27	0.18
Wheat Seed dataset	0.92	0.86	0.86

The Accuracy of both the datasets after application of 10 fold cross validation with classification into training and testing datasets is shown in the table below. We can see that the accuracy for iris dataset is lesser than that of seed dataset after cross validation.

TABLE 2: Accuracy of 10/5 - fold Cross validations

Data set	Accuracy
Iris dataset	0.27
Wheat Seed dataset	0.86

Moreover, the Accuracy of both the datasets after application of Rank-batch approached of Active learning with classification into training and testing datasets is shown in the below table. Here we were suprised by the outcome as their is significant difference noticed in the dataset after application of active learning rank algorithm. The Accuracy of iris dataset remarkably increased from 0.33 to 0.95. It had same impact on the seed dataset as the accuracy raised from 0.86 to 1.2 significantly.

TABLE 3: Accuracy before Rank-batch Mode

Data set	Accuracy
Iris dataset	0.33
Wheat Seed dataset	0.86

TABLE 4: Accuracy after Rank-batch Mode

Data set	Accuracy
Iris dataset	0.95
Wheat Seed dataset	1.0

The Accuracy of both the data sets after application of Pool based sampling approached of Active learning with classification into training and testing datasets is shown in the below table.The initial accuracy for both the datasets were calculated and the results wasn't satisfying as they are lowest at 0.33 and 0.86(Although it is quite good enough for seed data set).But after application of pooling the accuracy of Iris and Wheat seed data set has gradually increased at 0.88 and 1.2 respectively.

TABLE 5: Accuracy Before Pool Based Sampling

Data set	Accuracy
Iris dataset	0.33
Wheat Seed dataset	0.86

TABLE 6: Accuracy after Pool Based Sampling

Data set	Accuracy
Iris dataset	0.88
Wheat Seed dataset	1.2

6 Discussion

The model selection is as important as devising an active learning approach and choosing one classifier and one performance measure can often lead to unexpected and unwarranted conclusions.In real world application, active learning is a strategy which is not hard to deploy but is hard to perfect. It works best in the cases of unlabeled data such as news articles, tweets,images etc. But the labeling is expensive and will have the greatest impact on the model's training data to classify them in accurately labeled dataset.[5]

The Real-World Examples of active learning are as follows:
1.The Spam Filter Imagine a spam filter: its initial work at filtering email relies solely on machine learning. By itself, machine learning can achieve about 80–90% accuracy. The Accuracy of the filter is improvises when the user corrects the machine's output by labeling the exact messages that are not spam, and vice versa. Those relabeled messages feed back into the classifier's training data for finer tuning of future email. Another real-world example of active learning involves the ranking of online search results. Several years ago at Yahoo wanted to increase its ranking to top search results. The project involved identifying the

top 10 search results amongst millions. Thus, the top results, the classifier might assume that a machine-generated page repeatedly a thousand times is more relevant than another page with just a few mentions which is not necessarily the case.Hence, use of Active learning is significant.

Our task was to perform Active learning using cross validation on 2 dataset to check which method is the best fit in order to improve overall performance of the dataset. It is clearly seen that in the active learning phenomenon, we select some samples from the training data to the classifier in chunks to see the most informative samples and add them to the classifier.The same process was carried out in the K-fold cross-validation.As we split the data to k-subsets and train the model on the k-1 subsets leaving one for test/validation. So, the two algorithms have a common phenomenon of dividing the data into chunks and add them to the model. Active learning was introduced to handle the problem we data scientist have to face whilst having to work on label huge data.

However, among the other alternatives, we can label the datasets automatically using some pretrained algorithms, etc.The deep learning plays an important role in machine learning. Nevertheless, there's risk involved in mislabelling the data which might lead to type-1 and type-II errors.Therefore, Active learning was significantly used all over the globe as the best fit to handle large data sets where we can avoid labelling the entire data and can get more appropriate results. [6] However, we also need to keep in mind that during deployment, active learning is a one-shot problem and an evaluation set is not available. Hence, it is not possible to estimate the performance of the classification system during learning.Because it is difficult to decide when the system fulfilled the requirement and needs to terminate the work effectively.

Last but not the least, in active learning we train the model to achieve good performance in a greedy manner. We do not need to train with all the training data. Because in k-fold cross validation runs for the k-folds iterating through the entire data and consumes more time. Hence , this ability to achieve good performance and accuracy on small samples of the data is less time consuming and more liable.

7 Conclusion

In this paper, a demonstration have been made on how active learning and k-fold cross validation can be used to predict the target classes.Moreover, The paper discusses the impact of cross validation and active learning algorithms on the datasets. The study also supports the use of active learning in order to increase the model accuracy and increases the variance in the model.

A more thorough analysis could include that going through a new range of active learning algorithms such as rank-batch, pool based, Bayesian algorithm improves the accuracy and prediction rate of the model than that of old k-fold cross validation approach.The paper have discussed only 2 new structures and techniques of active learning out of many that are present in the real world. It would therefore be very exciting to see how proposed structure works when integrated into modern system learning techniques.

References

[1] A. Sologuren, "Active learning in machine learning," 2020. [Online]. Available: https://towardsdatascience.com/active-learning-in-machine-learning-525e61be16e5

[2] J. Prendki, "Introduction to active learning," 2018. [Online]. Available: https://www.kdnuggets.com/2018/10/introduction active-learning.html

[3] T. J. Paul, "Uci machine learning repository," 2010. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/seeds

[4] unknown, "Ranked batch-mode sampling," 2018. [Online]. Available: https://modal-python.readthedocs.io/en/latest/content/examples/

[5] T. Cuzillo, "Active learning vs k-fold cross-validation," 2015. [Online]. Available: https://www.oreilly.com/content/real-world-active-learning/

[6] S. eswar, “Active learning vs k-fold cross-validation,” 2019. [Online]. Available: https://medium.com/@sathiraju.eswar/active-learning-vs-k-fold-cross-validation-52b71ad8a181