Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Red wine quality prediction through active learning

To cite this article: Zhou Tingwei 2021 *J. Phys.: Conf. Ser.* **1966** 012021

View the article online for updates and enhancements.

Red wine quality prediction through active learning

Zhou Tingwei[*]

University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China

[*]Corresponding author's e-mail: 2018051408022std.uestc.edu.cn

Abstract. General machine learning requires a large number of training samples, meaning a lot of manpower and material resources. To solve this problem, active learning and its algorithms are often used. Active learning is a form of semi-supervised machine learning where the algorithm can choose which data it wants to learn from and then use the smallest and most effective labeled data set to make predictions or classifications. This article will show an example of using active learning to predict red wine quality. The predictive modeling approach the author chose was the K-Nearest Neighbor and the active learning algorithm was ranked batch-mode sampling. By observing the learning curve, the author found that generally the prediction accuracy of active learning would increase as the number of iterations increased. The author compared the experiment with another case using classic iris flower data set, and concluded that the prediction accuracy of active learning for different data sets depends on many factors, such as the correlation between the independent and dependant variables, the size of the data set and the number of iterations.

1. Introduction

Most machine learning research treats the learner as a passive recipient of data to be processed. This passive approach ignores the fact that, in many situations, the learner's most powerful tool is its ability to act, to gather data, and to influence the world it is trying to understand[1]. Therefore, a learning method that actively chooses the data it wants to learn is much more advanced and needed. Active learning, as a form of semi-supervised machine learning, has exactly the required features. With this approach, the program can actively query an authority source, either the programmer or a labeled data set, to learn the correct prediction for a given problem. This makes it possible to solve the dilemma that there are very few labeled samples in a large data set or it is very difficult to obtain labels. Strategies for querying the most informative instances are divided into six categories, and among all these strategies, uncertainty sampling is the simplest and most commonly used framework in which the labels of instances about which an active learner feels most uncertain are queried[2]. This paper will show an example of performing classification using an active learning method that includes this query strategy.

The data set used in this experiment contains several important indicators of 1600 bottles of red wine, as well as the quality assessed by famous wine tasters (from 0 to 10 stars)[9]. The author combined the indicators like fixed acidity,chlorides and density into a feature matrix. Then the author labeled the ones with quality of not less than 7 as "good quality", and labeled the rest as "bad quality", and then regarded quality as the prediction result. In order to achieve a higher prediction accuracy, the author picked the K-Nearest Neighbor as predictive modeling approach and ranked batch-mode sampling as active learning algorithm. The prediction accuracy could reach nearly 90% after several iterations, but it was still lower than the accuracy of the classic active learning cases.

Therefore,the author compared the experiment with the similar experiment using famous iris data set.The similarity of these two experiments is that they both combine the independent variables in the data set into a feature matrix, and then let the dependent variable be the result to be predicted. The difference lies in the correlation between the independent variables and the dependent variables, the size of the data set and the number of iterations,etc. These parameters that will be different in different learning objectives are the factors that affect the accuracy of prediction.

2. Analysis

2.1. PCA transformation

Now the author will introduce the specific implementation process. First, upload the data set on jupyter notebook, convert the first 11 columns of the data set into a feature matrix, and regard the dependent variable i.e. quality as the target to be predicted. Then for visualization purposes, PCA (principal components analysis) can be applied to the original data set. PCA is a feature extraction method, which extracts high-dimensional data to new data with dimensions lower than the initial dimension[3]. It is a method to reduce the features of high dimension by transforming multiple variables that may be correlated into a set of linearly uncorrelated variables through orthogonal transformation. This preprocessing stage of classification process helps getting more visual information through fewer variables. After PCA transformation, the data set looks like figure1 shown below. The red dots represent the red wine with quality of not less than 7, meaning good quality, while the purple dots are the others with poor quality.

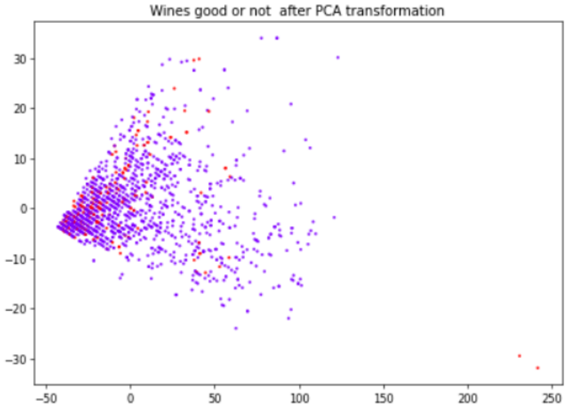


Figure 1 The visual classification diagram of wine quality after PCA transformation

2.2. Active learning method

The next step is to partition the data set into a training set and an unlabeled pool. First specify the training set consisting of 60 random examples. The remaining examples go to the “unlabeled” pool. Then apply the K-Nearest Neighbor classifier.

K-Nearest Neighbor algorithms have been widely used for classification and regression. These algorithms are based on searching the k points in a reference database that are closer to the measured data according to a function representing the distance between the two[4]. Then, the final solution is computed based on the k points that provide the minimum values of the distance. In the industry, K-Nearest Neighbor algorithms are more widely used in classification problems. The input consists of the k closest training examples in the feature space and the output is a class membership. Here if the author set the k value to 5, then the classification result of the unknown sample will be jointly determined by the five nearest labeled samples. When the author used the K-Nearest Neighbor classifier and the initial 60 training data to make prediction, the accuracy reached 80.4%.

Then the author used a method called ranked batch-mode sampling to update the model. This query strategy is an improved method of batch-mode active learning(BMAL) belonging to uncertainty sampling. The advantage of BMAL is that it does not need to retrain the model many times during a single selection step and is more suitable on some parallel labeling platforms[5]. Ranked batch-mode sampling reaches even better results by establishing a ranking among the batch. It relaxes traditional BMAL methods by generating a query whose answer is an optimized ranked list of instances to be labeled, according to some quality criteria, allowing batches to be of arbitrarily large sizes[6].

2.3. The process of active learning

The author allowed the classifier to query at most 450 instances it hadn't seen before.The batch size is 30, meaning there were 15 times of iteration, and each time it would balance the ideas behind uncertainty and dissimilarity in its choices. Figure 2-2 shows the prediction result by active learner using just original 60 training data. The green pluses represent the samples with correct predictions, and the red crosses represent the samples with wrong predictions. It is not difficult to find that the samples with incorrect predictions are concentrated in the two-dimensional image, indicating that samples with certain similar characteristics are easy to be predicted incorrectly. And the areas where the red crosses are concentrated are the very areas where red wines of good quality are concentrated and areas nearby.

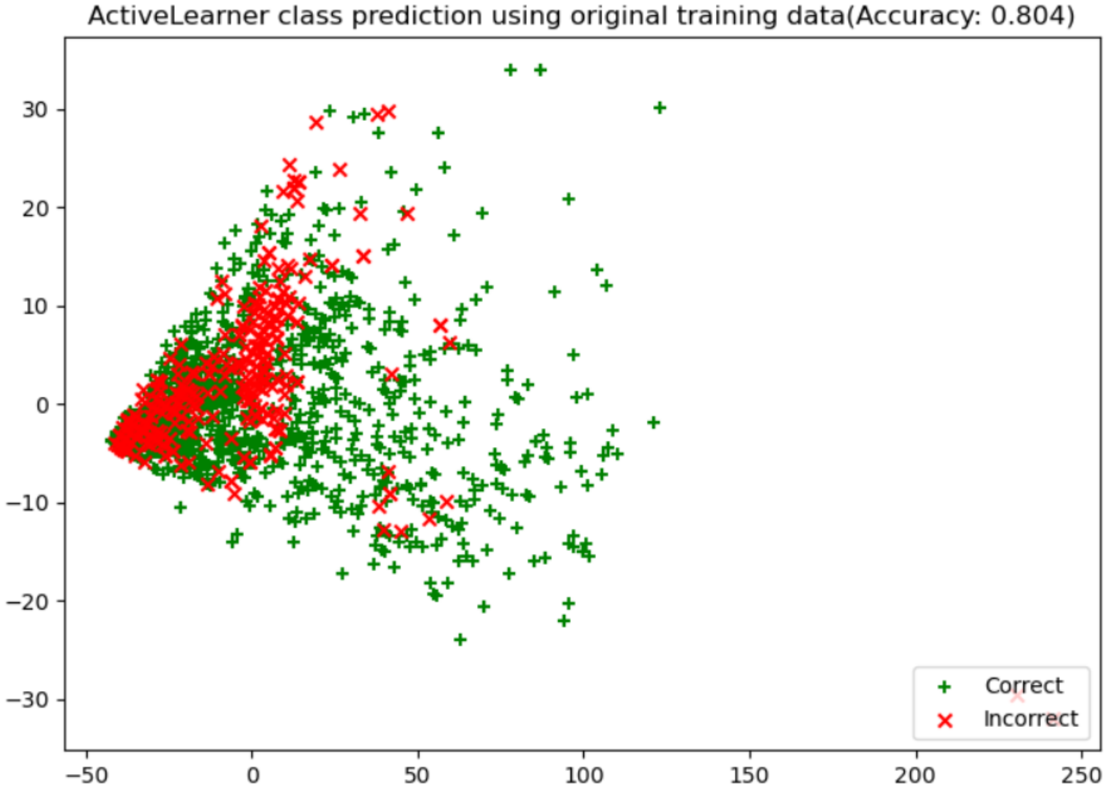
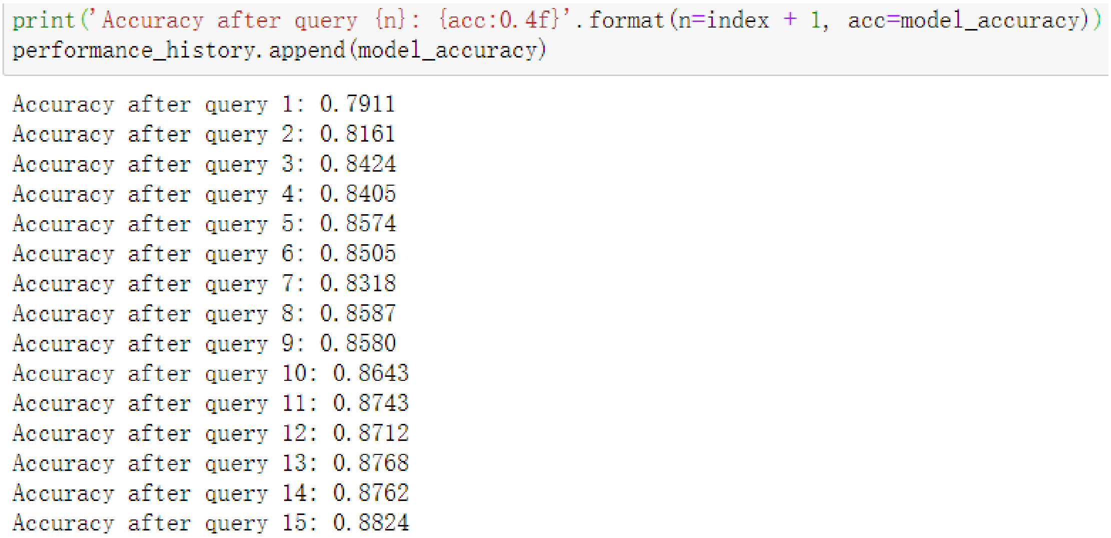


Figure 2:The prediction result of wine quality by active learner using original training data

The values of prediction accuracy after each iteration are recorded in figure 3. Figure 4 is a line graph that reflects the complete learning curve of this active learning. It can be seen from figure 4 that the learning curve generally conforms to the characteristics of active learning, that is, the prediction accuracy first increases faster as the number of iterations increases, and then slowly increases to a stable point. However, there are some abnormal pits indicating that the prediction accuracy does not strictly increase with the number of iterations. From this, a conjecture can be put forward. The prediction accuracy does not increase with the increase of the number of samples that have been learned. Instead, samples of different classifications with similar characteristics may become feedback that biases the classifier temporarily.



```
print('Accuracy after query {n}: {acc:0.4f}'.format(n=index + 1, acc=model_accuracy))
performance_history.append(model_accuracy)
```

```
Accuracy after query 1: 0.7911
Accuracy after query 2: 0.8161
Accuracy after query 3: 0.8424
Accuracy after query 4: 0.8405
Accuracy after query 5: 0.8574
Accuracy after query 6: 0.8505
Accuracy after query 7: 0.8318
Accuracy after query 8: 0.8587
Accuracy after query 9: 0.8580
Accuracy after query 10: 0.8643
Accuracy after query 11: 0.8743
Accuracy after query 12: 0.8712
Accuracy after query 13: 0.8768
Accuracy after query 14: 0.8762
Accuracy after query 15: 0.8824
```

Figure 3. The values of prediction accuracy of red wine quality after each iteration

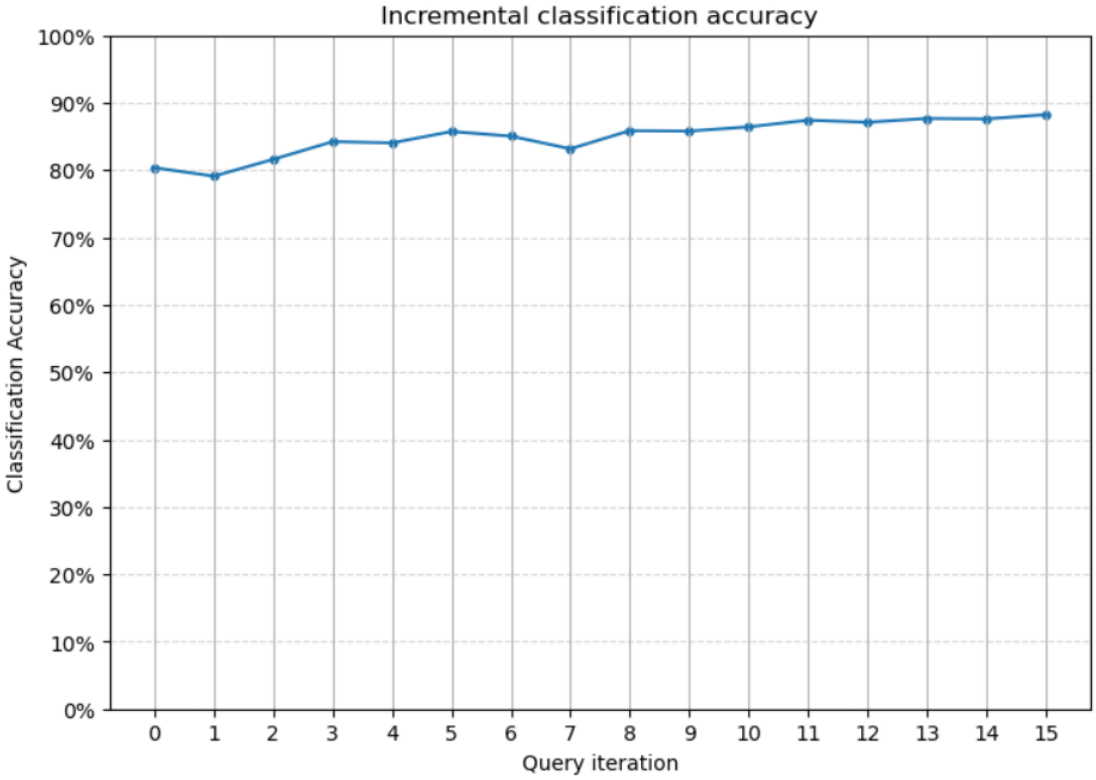


Figure 4. The learning curve of red wine quality using ranked batch-mode sampling

Figure 5 shows the prediction result after 15 queries. Compare it to figure 2, and it can be found that those areas that are still occupied by red crosses are almost unchanged, but have become smaller. This shows that some samples in this data set do have similar characteristics to samples of another classification, and therefore are always predicted incorrectly.

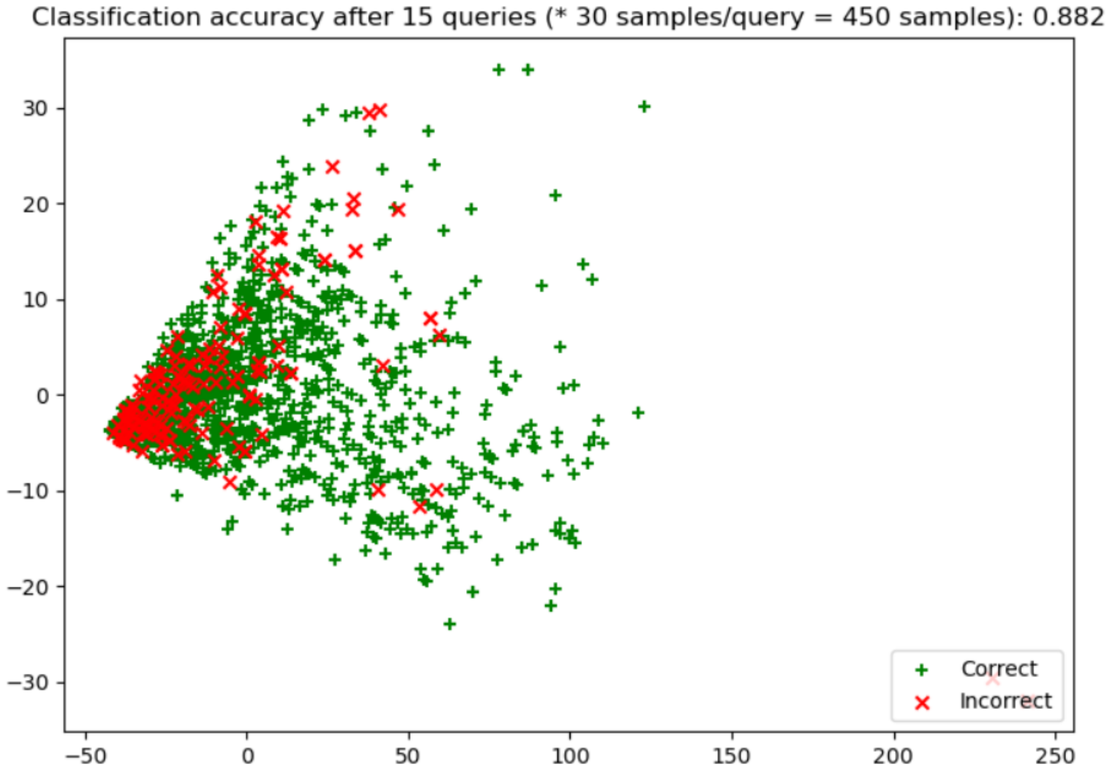


Figure 5. The prediction result of red wine quality by active learner after 15 queries

3. Comparison

The most famous data set for active learning is the iris flower data set. There are three different kinds of iris flowers in this data set, each with 50 samples. Previously, researchers have combined some characteristics of iris flowers like sepal length and petal width into a feature matrix and regarded the kind of iris flowers as the label to conduct an active learning research.They picked the same predictive modeling approach and active learning algorithm as the author picked, and they obtained a high value of prediction accuracy. The visual classification diagram of iris flowers after PCA transformation is shown in figure 6. The three different kinds of colored dots represent three kinds of iris flowers, each occupying an area in the two-dimensional image, and these areas hardly intersect. Then look at the two-dimensional image of red wine quality, it can be seen that the red wines of better quality are scattered in the rest red wine, and there is no obvious boundary. This is one of the reasons why the prediction accuracy of red wine quality can not reach the level of prediction of iris flower.

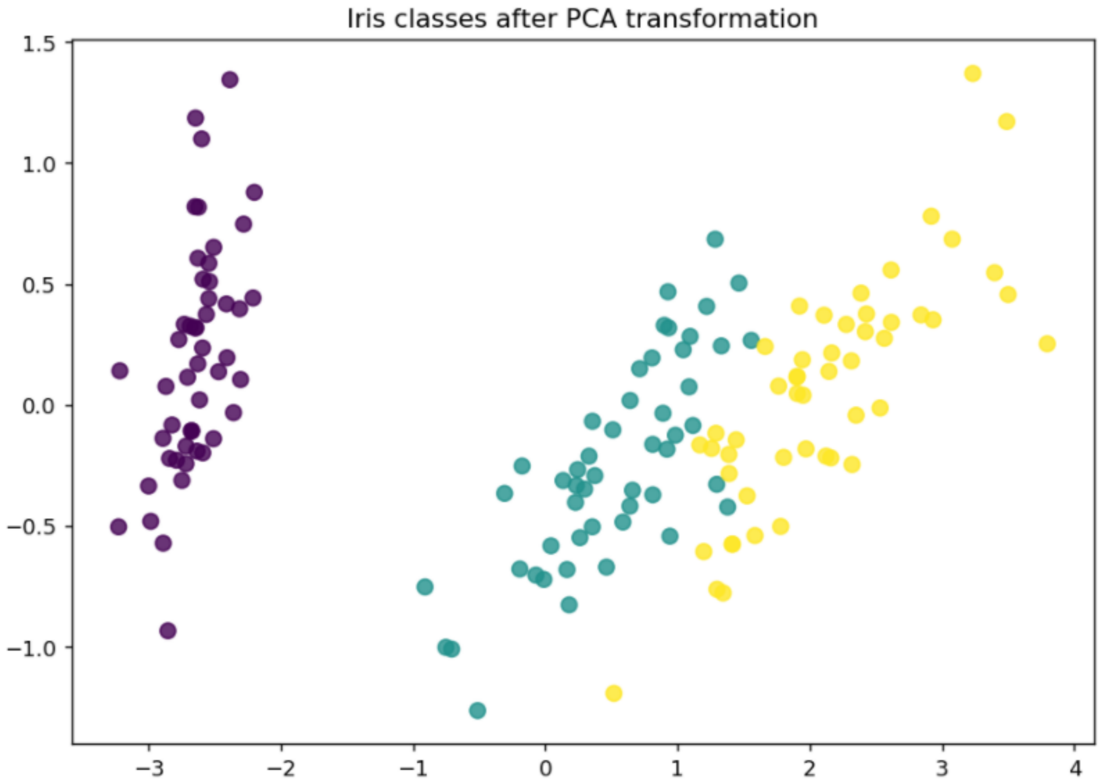


Figure 6. The visual classification diagram of iris classes after PCA transformation

The learning curve of iris flower prediction and the prediction result by active learner are shown below in figure 7 and figure 8. The iris data set can reach a higher prediction accuracy after fewer iterations. The prediction accuracy is approximately 95% after 5 times of iteration. Compare these two experiments, and it can be found that the difference lies in the correlation between the independent variables and the dependent variable, the size of the data set and the number of iterations. Each kind of iris flowers has a general range of sepal length and petal width. When several characteristics of an iris flower satisfy a certain kind of iris flower, the probability to get correct prediction result is high. However, the quality of red wines was provided by the tasters or the experts. As we all know, wine tasting involves precise sensory evaluations, individual characteristics inherent in each expert along with their physical and psycho-emotional state also contribute to the subjectivity of expert evaluation[7]. The subjectivity of this judgment will also cause the outcome of the dependent variable to be less predictable. Moreover, according to the provider of the data set, some relevant data was not provided due to the interest issues, which means the independent variables that affect the dependent variable are not complete. Apart form that, the data set of red wines is much bigger than the data set of iris flowers, and the number of indicators for red wine as well as the ranges of indicator values are much larger than the iris data set. Therefore, it is unrealistic for the red wine quality experiment to learn just a few samples to achieve high prediction accuracy like the iris flower experiment.

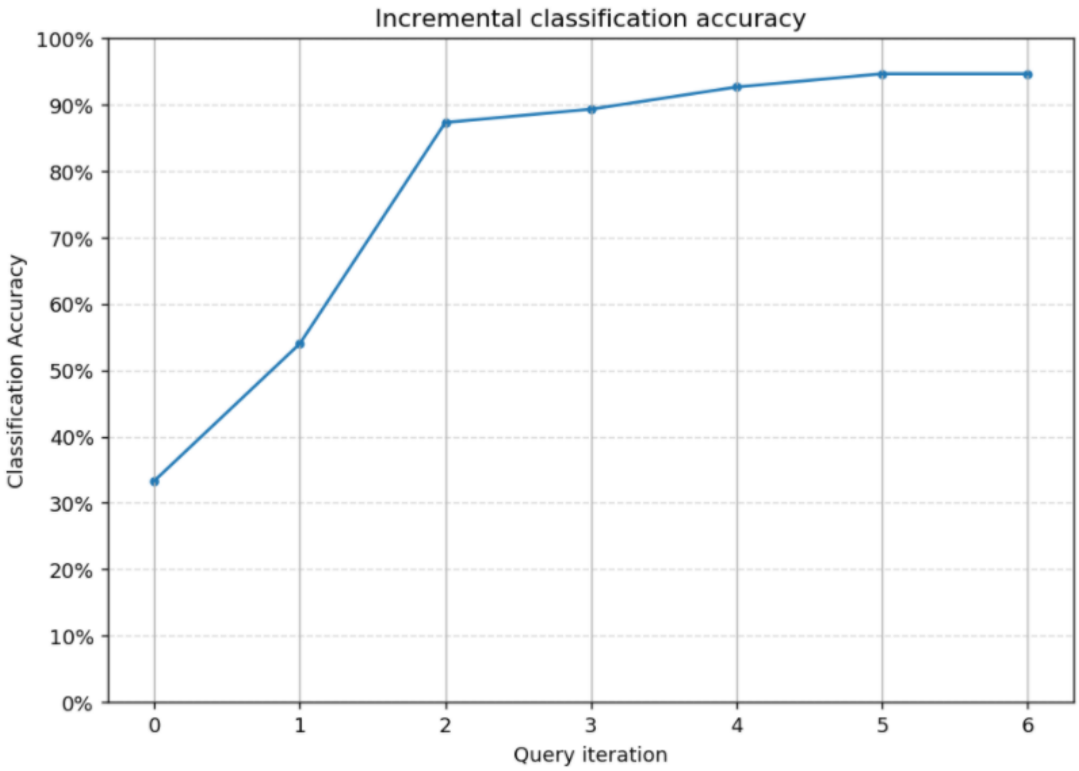


Figure 7. The learning curve of iris classes using ranked batch-mode sampling

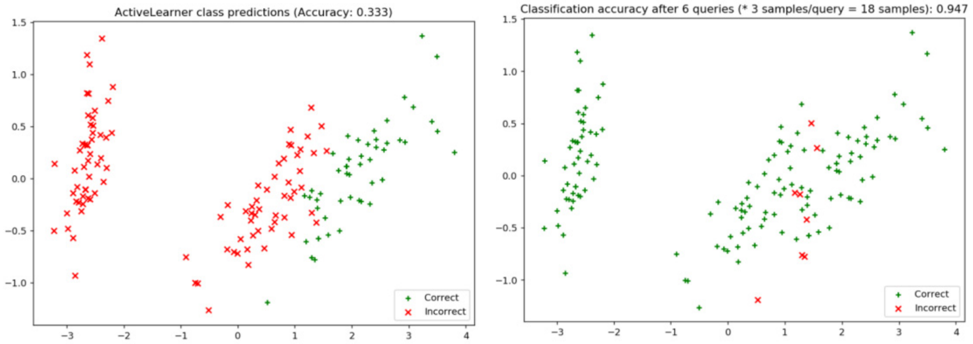


Figure 8. The prediction result of iris classes by active learner(a)using original training data(b)after 6 queries

4. Suggestion

When using active learning to predict different data sets, it should be noted that there are three broad scenarios and many specific query strategies. The success of active learning hinges significantly on the query strategies, i.e., how to select observations for user feedback. A poor choice can waste scarce resources, like time and intellectual effort of human annotators, without achieving any improvement[8]. General query strategies include least confidence, margin sampling and entropy sampling, etc. The query strategy used in this article is about ranking the batches of samples and deciding one batch for the active learner to learn according to a selection criterion. It should be emphasized that this query strategy may not achieve the best result for this learning objective. If for different learning objectives, the corresponding most suitable active learning algorithm or query strategy is known, the probability of failed experiments can be reduced and the effect of experiments can be improved.

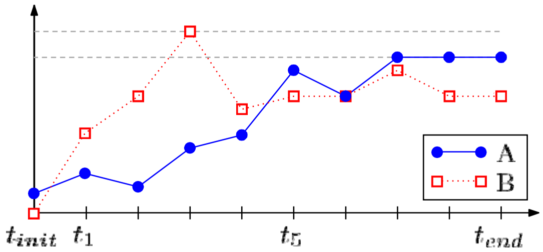


Figure9. The learning curves for active learning methods A and B

Figure 9 shows the comparison of two learning curves that are obtained through different active learning methods A and B under the same learning objective. The y-axis is the value indicating the prediction accuracy and the x-axis is the value indicating the number of learning iterations. It can be seen that better results can be obtained in the first few iterations of method B, but after a certain peak, the y value will decrease with the number of iterations. After that, even when y value increases again, it will not reach the peak height. In method A, the y value generally increases with the number of iterations, and when it finally reaches a stable value, it is higher than that in method B. Therefore, when researchers hope to achieve higher prediction accuracy with fewer label samples, they can choose method B, while if the target data set has many label samples, and they hope to use more iterations to get the highest and reliable prediction accuracy, then they can choose method A. The experiment in this article requires a large number of active learning iterations and the prediction accuracy generally shows a positive increase, so it is similar to the situation of using method A.

In fact, researchers have done related experiments before. They introduced 8,400 active learning methods that are composed of different learning scenarios, active learning classifiers and query strategies to operate on a learning objective. Using the learning progress curve as an evaluation criterion, they came to a conclusion: none of the state-of-the-art methods stands out in a competitive evaluation. The performance largely depends on the parametrization of the classifier, the data set, and on how progress curves are summarized[8]. This reflects the correctness of the no-free- lunch theorem.

Therefore, subsequent researches on active learning can focus on finding a formula with better generality. When the size of the data set, the number of iterations required and the correlation between the independent variables and the dependent variable are parameterized and then substituted into the formula, the best applicable active learning methods can be obtained. When this is implemented, active learning will be more universally applied to various classification problems.

5. Conclusion

This article explains in detail how to predict the quality of red wines by an active learning method combined with K-Nearest Neighbor algorithm and ranked batch-mode sampling. After being trained by the initial 60 training data and querying for 15 times of iteration, the active learner achieved a prediction accuracy of 88.2%. Compared to the classic iris flower experiment, the prediction accuracy of red wine quality is a bit lower. This is due to the low correlation between the independent variables and the dependent variable, the incompleteness of the data set and overly complex indicators.

Although in general this experiment predicted a highly subjective data set and still achieved a high prediction accuracy, the problem still lies in that it required a lot of labeled data. The initial training data is only 60, but as the number of iterations increases, the total labeled data obtained exceeds 30% of the total number of samples. As the author mentioned in the analysis of figure 2-4 before, the increase in prediction accuracy with the increase in the number of iterations is not obvious, and there are even negative increases at some points. This will weaken the persuasiveness of the advantages of active learning to a certain extent. Therefore, what the author should do next is to collect more index data that affect red wine quality, integrate them into a larger data set and constantly modify some important parameters such as batch size and number of iterations. Then the most important factors that affect the prediction accuracy of this active learning method should be explored.

Acknowledgments
First of all, the author would like to express sincere gratitude to the professor and teaching assistant for their patient guidance and assistance in the process of comprehending the research content from the shallower to the deeper and finally completing this thesis. At the same time, the author would also like to thank all the students who have the same research direction and enthusiasm for their support and encouragement. Without their enlightening guidance and patient company, the author cannot finish this thesis.

References

[1] Cohn D. A.,Ghahramani Z.,Jordan M. I.. Active Learning with Statistical Models[J]. Journal of Artificial Intelligence Research,1996,4.

[2] Yong Zhang, Meng Joo Er, Sequential active learning using meta-cognitive extreme learning machine, Neurocomputing, Volume 173, Part 3, 2016, Pages 835-844, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2015.08.037.

[3] R Fauzi et al 2021. Journal of Physics: Conference Series, Volume 1725, The 2nd Basic and Applied Sciences Interdisciplinary Conference 2018 (2nd BASIC 2018) 3-4 August 2018, Depok, Indonesia

[4] Kun Feng, Arturo González, Miguel Casero, A kNN algorithm for locating and quantifying stiffness loss in a bridge from the forced vibration due to a truck crossing at low speed, Mechanical Systems and Signal Processing, Volume 154, 2021, 107599, ISSN 0888-3270, https://doi.org/10.1016/j.ymssp.2020.107599.

[5] Y. Yang, X. Yin, Y. Zhao, J. Lei, W. Li and Z. Shu, "Batch Mode Active Learning Based on Multi-Set Clustering," in IEEE Access, vol. 9, pp. 51452-51463, 2021, doi: 10.1109/ACCESS.2021.3053003.

[6] Thiago N.C. Cardoso, Rodrigo M. Silva, Sérgio Canuto, Mirella M. Moro, Marcos A. Gonçalves, Ranked batch-mode active learning, Information Sciences, Volume 379, 2017, Pages 313-337, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2016.10.037.

[7] Alexan A. Khalafyan, Zaual A. Temerdashev, Vera A. Akin'shina, Yuri F. Yakuba, Data on the sensory evaluation of the dry red and white wines quality obtained by traditional technologies from European and hybrid grape varieties in the Krasnodar Territory, Russia, Data in Brief, Volume 36, 2021, 106992, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2021.106992.

[8] Holger Trittenbach, Adrian Englhardt, Klemens Böhm, An overview and a benchmark of active learning for outlier detection with one-class classifiers, Expert Systems with Applications, Volume 168, 2021, 114372, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.114372.

[9] Ranked batch-mode sampling, 2018, https://modal-python.readthedocs.io/en/latest/content/examples/ranked_batch_mode.html