Review on Evaluation Criteria of Machine Learning Based on Big Data

HaoWu¹, FanJunMeng¹

¹Inner Mongolia Normal University, Hohhot, Inner Mongolia, 010020, China

Abstract. In machine learning in the context of big data, it is an indispensable link for the evaluation results and indicators of prediction results. Although it is not as important as the core algorithm, it plays a role in related applications and research fields. The role of horizontal comparison of learning algorithms. The article briefly summarizes and reviews the research results of basic machine learning algorithm evaluation standards. Different evaluation methods are applied to different machine learning problems, and the trends and difficulties of machine learning evaluation methods in the future are analyzed.

1. Introduction

In the conditions of big data, along with the rapid artificial intelligence’s development, especially on the Internet of rapid development, in-vehicle mobile terminals, and cloud technologies, big data is generally reflected in many environments. According to a recent Dell company survey, it has reached 2.8ZB (1ZB = 1021B); IBM survey shows that the total amount of data in the world generates 25 billion bytes of big data every day, it is estimated that it may reach 40ZB next year, equivalent to more than 300 times in the last 15 years. Nowadays, big data technology has been applied to most fields, including transportation, energy, service industry, and medical care. It has received extensive attention from many companies, universities, research institutes, and governments.

There are many definitions of machine learning, including the following: "Machine learning is a direction that focuses on the improvement of the system's performance through scientific computing and the use of the experience gained." "Machine learning is derived from artificial intelligence and statistics" "Machine learning is the study of how computers imitate human learning behaviours, acquire new knowledge, and improve their performance." Existing machine learning methods have been very mature and improved. For most classification, regression, and clustering problems, they can be roughly divided into semi-supervised learning, unsupervised learning, supervised learning, reinforcement learning, and algorithms. The development has been divided into algorithms, regularization methods, decision tree learning, instance-based algorithms, Bayesian algorithms, clustering algorithms, association rule learning, genetic algorithms, artificial neural networks, kernel-based algorithms, integrated algorithms, etc. However, there are few types of research on evaluation methods based on machine learning under big data at home and abroad. Therefore, it is not yet possible to verify whether a learning algorithm is suitable for a big data environment. Influences. Foreign scholars have proposed in the literature to evaluate machine learning algorithms using three criteria: recall rate, Matthews correlation coefficient, and accuracy. Big data environment and unsupervised, weakly supervised machine learning algorithms for evaluation. This article summarizes the existing evaluation methods used by big data machine learning to solve most problems and summarizes the best evaluation methods for different problems to fill the gaps in the evaluation methods of big data machine learning algorithms.
2. Basic Evaluation Criteria Commonly Used in Machine Learning

In the big data environment, how to evaluate the pros and cons of a machine learning algorithm becomes more and more important, because a good evaluation system is more instructive for machine learning algorithms oriented to big data.

2.1 Accuracy and Recall

For the binary classification problem, the examples can be divided into: True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN). You can then build the Confusion Matrix as shown in the following table 1.

| Real category | Forecast category |
|---------------|------------------|
| Positive example | TP | FN |
| Negative example | FP | TN |

(1) Accuracy, $A$

$$A = \frac{TP}{TP + FP}$$

(2) Precision, $Pr$

$$Pr = \frac{TP}{TP + FP}$$

(3) Recall, $R$

$$R = \frac{TP}{TP + FN}$$

(4) F1-score, $F_\beta$

$$F_\beta = \frac{(1 + \beta^2) \times Pr \times R}{(\beta^2 \times Pr) + R}$$

Also, the G-Mean index can also evaluate the model performance of imbalanced data.

The accuracy is used to evaluate the prediction ability of the model as a whole, but the gap between the amount of positive and negative samples of the training sample is too obvious, which may cause the model to have a low accuracy (low recall) or too strong (low recall) of positive samples when the model has high accuracy Precision is low. The accuracy rate is often used in the problem of binary classification. Therefore, the recall rate and accuracy rate need to be specifically assess the excellence of the model. In [9], the accuracy of the prediction in the experiment is to predict how many samples of poor students are really poor students the higher the accuracy, the better the ability to identify a negative example. For multi-classification problems, it can be transformed into two-class classification problems for each category to measure model quality. Retrieval and selection of large-scale data collections are related to the two indexes of "recall rate" and "accuracy rate". And because the two indicators are mutually restrictive, we usually choose a suitable degree for the "retrieval strategy" as needed, not too strict or too loose, Find a balance between recall and accuracy, this balance is determined by specific needs.

2.2 ROC and AUC

ROC is expressed as receiver-operating characteristic. Each point on the ROC curve corresponds to the susceptibility of the same signal. In the ROC curve, the horizontal axis represents the false positive rate (FPR), also called specificity. The proportion of negative examples; (1-Specificity), where the vertical axis represents the true positive rate (TPR), also called Sensitivity. For a binary classification problem, the instances are classified into positive or negative. But in practice, there are four situations.

(1) Assuming that the instance is a positive class, it is predicted to become a positive class, thatis, the true class (True Positive TP)
(2) Suppose the instance is a positive class and is predicted to become a negative class, that is, a false negative class (False Negative FN).

(3) Assuming that the instance is a negative class and is predicted to become a positive class, it is a false positive class (False Positive FP).

(4) Suppose the instance is a negative class and is predicted to become a negative class, which is a true negative class (True Negative TN).

TP: correct positive number

FN: underreported, no number of correct matches found

FP: False positive, no match is incorrect

TN: Number of mismatches rejected correctly

As shown in the table 2, 1 represents the positive class, and 0 represents the negative class:

| Actual | 1 | 0 | Total |
|--------|---|---|-------|
| 1      | True Positive TP | True Negative FN | Actual Positive(TP+FN) |
| 0      | False Positive FP | True Negative TN | Actual Negative (FP+TN) |
| Total  | Predicted Positive(TP+FP) | Predicted Negative(FN+TN) | TP+FN+FP+TN |

From this, we can get the formula

1. True Postive Rate (TPR): TP / (TP + FN), which represents the proportion of actual positive instances to all positive instances in the positive class predicted by the classifier. Sensitivity

2. False Postive Rate (FPR): FP / (FP + TN), which represents the proportion of actual negative instances to all negative instances in the positive class predicted by the classifier. 1-Specificity

3. True Negative Rate (TNR): TN / (FP + TN), which represents the ratio of actual negative instances to all negative instances in the negative class predicted by the classifier, TNR = 1-FPR. Specificity

Suppose a logistic regression classifier is used, which gives the probability of being a positive class for each instance, then by setting a threshold value such as 0.8, the possibility is greater than or equal to 0.8 as the positive class, and the fewer than 0.8 is the negative class. Correspondingly, a group (FPR, TPR) can be calculated, and corresponding coordinate points can be obtained in the plane. As the threshold value gradually decreases, more and more instances are classified as positive classes, but these positive classes are also doped with true negative instances, that is, TPR and FPR will increase at the same time. When the threshold is maximum, the corresponding coordinate point is (0,0), and when the threshold is minimum, the corresponding coordinate point is (1,1). As shown in figure 1.

![Figure 1. ROC and AUC.](image)

Horizontal axis FPR: 1-TNR, 1-Specificity, the larger the FPR, the more actual negative classes are predicted in the positive class.

Vertical axis TPR: Sensitivity (positive type coverage), the larger the TPR, the more positive types are predicted in the positive type.
Ideal target: TPR = 1, FPR = 0, which is the (0,1) point in the figure, so the closer the ROC curve is to the (0,1) point, the better of the 45 degrees diagonal, the greater the Sensitivity and Specificity.

The ROC curve has a very good characteristic: when the distribution of the positive and negative samples in the test set is transformed, the ROC curve can remain unchanged. In actual data sets, sample imbalances often occur, that is, the difference between the positive and negative samples is large, and the positive and negative samples in the test data may also change over time. The following figure is a comparison of ROC curve and Precision-Recall curve, As shown in the graph.

![Figure 2: ROC Curves.](image)

In figure 2 above, the two on the left are ROC are ROC curves, and the two on the right are ROC Precision-Recall curves.

(a) And (b) show the results of classifying the first test set (the distribution of the positive and negative samples is balanced), and (c) and (d) are the classifiers that increase the amount of negative samples in the test set by 10 times the first. As a result, it can be clearly seen that the ROC curve basically maintains the original appearance, while the Precision-Recall curve changes greatly.

3. Other Basic Evaluation Criteria for Machine Learning

3.1 Kappa Value

The kappa value is a measure of classification accuracy. KIA is mainly used to compare and analyze whether the difference between two maps or images is caused by “accidental” or “inevitable” factors. It is also often used to check the correctness of satellite image classification for the determination of real features. KIA is an index capable of calculating overall consistency and classification consistency. The general expression of KIA’s calculation formula is:

\[
K = \frac{\text{observed accuracy} - \text{chance agreement}}{1 - \text{chance agreement}}
\]

In Literature[23], the detection of student performance prediction behaviour was found to be more intuitive than other evaluation methods, and a positive correlation was clearly obtained after the use of key indicators. The prediction results for the detection of malicious Android malware in Literature[24] used Kappa values to fully evaluate the performance of the classifier and obtained significant effectiveness and accuracy.
3.2 Gini Coefficient
When using SAS or some other statistical analysis software to evaluate the classification effect of the classifier, you often see the Gini coefficient. The Gini coefficient is often used to justice the fairness of income distribution.

Information gain is often used to choose features in the ID3 Algorithm, and preference is given to large information gain. Information gain ratio is used to choose features to reduce the question that data gain is easy to choose features with a lot of feature values in the C4.5 algorithm. However whether it is ID3, C4.5 or other, it is based on the entropy model of information theory, which involves a large number of logarithmic operations. Can you simplify the model without losing the advantages of the entropy model completely? The information gain ratio has been replaced in the CART classification tree algorithm by the Gini coefficient, where the Gini coefficient represents the model's impureness. The much higher the Gini coefficient, the higher the impureness and the worse the characteristics. It's the opposite of information gain.

Gini coefficient refers to the ratio of the section enclosed by line of fairness and Lorenz Curve to the area below line of fairness, that is, the Gini coefficient = A area / (A area + B area). ReferenceLiterature[28] used this evaluation method to have concrete manifestations, which clearly showed that for different vehicle trip sharing rates, this evaluation method is used to evaluate the prediction effectiveness of classification models, which refers to the section enclosed by this ROC curve and the centre line Proportion to area above midline.

4. Conclusions and Discussions
Machine learning algorithms under big data have always been research hotspots, and how to evaluate the efficiency and goodness of existing and new machine learning algorithms has also become very important. This article is based on a preliminary basic summary of machine learning model algorithms that are now mature. For the evaluation method of machine learning algorithms under big data, the basic process of rating method and the advantages of specific application environment are briefly described. The evaluation method summarized in this article is only a simple analysis of evaluation methods of machine learning algorithms, such as accuracy and recall. It can measure the accuracy of the binary classification algorithm, but it does not further research the comparison of different characteristics of the evaluation criteria of machine learning, and cannot research the pros and cons of the evaluation criteria in machine learning. Therefore, in future research, we will further research the pros and cons of different evaluation standards, and put forward specific use advantages such as evaluation standards for specific problems.

Acknowledgements
This research was financially supported by National Natural Science Foundation of China (Research on Construction of Big Data-driven Multi-dimensional Blended Teaching Model in Mongolian-
Chinese Bilingual SPOC Environment Gant No.61841703), Natural Science Foundation of Inner Mongolia of China (Research on Identification Algorithm of Genomic Structure Variation Based on Deep Neural Network in Cloud Computing Environment Grant No.2019BS06001), Natural Science Foundation of Inner Mongolia of China (Research on Application of Big Data Analysis Technology in Mongolian-Chinese Bilingual SPOC Environment Multidimensional Mixed Teaching Model Grant No.2019MS06014), Inner Mongolia Normal University Research start-up project (Grant No.2017YJRC020).

References
[1] Zhongjian Li, Xianmin Liu. An important aspect of big data: data availability [J]. Computer research and development, 2013, 50(6): 1147—1162. (In Chinese)
[2] Lei Deng. Hotel Price Prediction Analysis Based on Machine Learning [D]. Southeast University, 2017. (In Chinese)
[3] Jue Wang, Zhihua Zhou, Aoying Zhou. Machine Learning and Its Applications [M]. Tsinghua University Press, 2006. (In Chinese)
[4] Jianfeng Yang, Peirui Qiao, Yongmei Li, Ning Wang. Review of Machine Learning Classification Problems and Algorithms [J]. Statistics and Decision, 2019, 35(06): 36-40. (In Chinese)
[5] Panhao Zhen. Talking About Machine Learning Methods [J]. Network Security Technology and Application, 2014, 1: 176—177. (In Chinese)
[6] Smusz S, Kurczab R, Bojarski A J. Multidimensional Analysis of Machine Learning Methods Performance in the classification of bioactive compounds [J]. Chemometrics and Intelligent Laboratory Systems, 2013, 128: 89—100.
[7] Yongrong Zheng, Jiazheng Yuan, Hongzhe Liu, Chao Li. Evaluation method of machine learning algorithm for big data [J]. Computer science, 2015, 42(10): 116—119. (In Chinese)
[8] Peixin Yang. Research on distributed generation grid protection based on machine learning [D]. 2016. (In Chinese)
[9] Guiming Lu, Yuan Zhang, Zhimin Zhou. Research on classification and prediction of poor students based on machine learning [J]. Computer applications and software, 2019, 36(01): 316–319. (In Chinese)
[10] Xinyan Zhang. Research on Set Multi-label Text Classification [D]. Liaoning Technical University. (In Chinese)
[11] Ming Zhao. Research on Image Retrieval Methods Based on Machine Learning [D]. Dalian Maritime University, 2012. (In Chinese)
[12] Zhihua Zhou. Machine Learning [M]. Tsinghua University Press, 2016. (In Chinese)
[13] Yi Yao, Yang Li, Zhongyu Zhou, et al. Method for detecting malicious behavior of air ticket booking [J]. Electronic test, 2018, No.396(15): 45–47. (In Chinese)
[14] Zhijie Liu. Research on Melanoma Recognition Algorithm Based on Dermatoscope [D]. (In Chinese)
[15] Chenyu Ji. Research on Imbalanced Data Classification and Its Application in Bank Marketing [J]. Shanxi Electronic Technology, 2018, 200(05): 57–59+71. (In Chinese)
[16] Yanyun Wang, Songcan Chen. Review of AUC-based Classifier Evaluation and Design [J]. Pattern recognition and artificial intelligence, 2011, 24(01): 64–71. (In Chinese)
[17] Ling C X, Huang Jin, Zhang H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms // Proc of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence. Halifax, Canada, 2003: 329–341
[18] Ling C X, Huang Jin, Zhang H. AUC: A Statistically Consistent and More Discriminating Measure Than Accuracy // Proc of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003: 519–526
[19] Rosset S. Model Selection via the AUC // Proc of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 89–97
[20] Zhao Ding. Research and Application of Collaborative Learning Algorithms [D]., 2017. (In Chinese)
[21] Shuai Jiang. Research on Classifier Performance Evaluation Based on AUC [D]. Jilin University, 2016. (In Chinese)
[22] Lina Rong. Research on Change Detection of High Resolution Remote Sensing Image Based on Object-oriented Classification [D]. 2017. (In Chinese)
[23] Huan Liu. Research and Application of Performance Prediction Model Based on Student Behavior [D]. Xidian University, 2017. (In Chinese)
[24] Haijian Zhang. Android malicious application detection system based on deep learning [J]. Electronic Technology and Software Engineering, 2017(03): 227. (In Chinese)
[25] Wenhong Jia. "School Choice Fever" and Educational Gini Coefficient: Taking Primary and Secondary Schools as Examples [J]. WeiShi, 2007(11): 54-56. (In Chinese)
[26] Jing Hu. Research on User Network Behavior Recognition Technology Based on 3G Core Network Data [D]. (In Chinese)
[27] Aixiang He. Research on the Application of Decision Tree in Employment Management of Applied Undergraduate Universities [D]. 2011. (In Chinese)
[28] Xiaofeng Qi, Fang Cheng, Yupeng Zhang. Bilevel optimization model of public transport network based on group fairness difference [J]. Journal of China Highway, 2014, 27(10). (In Chinese)