Reanalysis of Classification Algorithms on Different Datasets

Peng-fei Yue[1a], Qin-ge Wu[1,b*], Jian-gang Zhu[2c], Wen-fang Cheng[2d],and Xiao-liang Qian[1f]

[1]College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

[2]Polar Research Institute of China, Shanghai, 200136, China

* Corresponding author

wqe969699@163.com[b]

Keywords: Classification, Data Mining, Analysis, Logistic Regression

Abstract: In this paper, we use four classification algorithms like C4.5, naive bayes, logistic regression and K nearest neighbor. It outlines basic principles of classification algorithms and we analysis and summary each algorithm has its advantages, disadvantages through performance comparison on different datasets.

Introduction

Classification is a classical problem in data mining such as retail, insurance[1].*C4.5 decision tree* has been widely used and also has many improved algorithms. Juan[3] use post-pruning C4.5 algorithm, it can solve geographical constraints and improve enrollment rate. *Naive bayes* is derived from classical mathematical theory. Sudhakar[4] propose a polynomial naive bayes algorithm for extract semantic relations of biomedical science, it can provide a understanding for patient's condition. *Logistic regression* model is easy to understand, especially in large scale linear classification. Ganser[5] use L2 regularized logistic regression algorithm, it reduces cost of medical care and improve health care situation. *K nearest neighbor* is an inert machine learning algorithm. For example, Pu[6] propose a parallel pipeline structure and bubble sort using FPGA method to optimize KNN algorithm, it is higher energy consumption than traditional KNN.

In this paper, we will make a general description of the four classification algorithms, and then compare performance on different datasets, such as accuracy, recall, precision. Finally, we give advantages, disadvantages.

Classification Algorithms

C4.5 Algorithm

C4.5 algorithm uses information gain rate to split dataset. It can not only deal with discrete attributes but also process continuous attributes. By pruning, it can avoid over-fitting phenomenon in some extent [2].

$$Ent(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

$$Gain(D, A) = Ent(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} Ent(D_v) \quad (2)$$

$$Gain\_ratio(D, A) = \frac{Gain(D, A)}{Ent(D)} \quad (3)$$

Naive Bayes Algorithm

Bayes is a kind of algorithm which uses knowledge of probability and statistics to classify, it mainly include naive bayes algorithm and bayes network algorithm. Naive bayes based on bayes theorem, that is, when estimated class conditional probability, attributes are independent of each other[2].

Logistic Regression Algorithm

Logistic regression algorithm is a nonlinear regression model. Feature can be continuous, classified variables or dummy variables, and it can predict probability of future results through historical data[2]. Logistic regression is classified by introducing Sigmoid function which is different from linear regression.

K Nearest Neighbor Algorithm

K nearest neighbor is a kind of common supervised learning classification algorithm. In short, K-nearest neighbor algorithm is used to measure distance between different feature values[2]. Common distance metrics are shown below:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}\ . \tag{4}$$

Experiment and Analysis

Datasets

There are four datasets we have used in our paper taken from UCI Machine Learning Repository[7] and Polar data sharing platform[8]. The details of each datasets are shown in Table 1.

Table 1. Details of four datasets

Datasets	Instances	Attributes	Class
trace metal	287	6	4
vehicle	946	19	4
adults	48842	14	2
ionosonde	100000	17	2

Analysis

This experiment uses accuracy, precision and recall to judge the performance of each algorithm. The graphs of performance indicators were plotted for C4.5, naive bayes, logistic regression and K nearest neighbor algorithm.

In Figure 1, K nearest neighbor has higher value. In Figure 2, logistic regression has higher accuracy than others. In Figure 3, C4.5 has higher accuracy than others. In Figure 4, naive bayes has better accuracy.

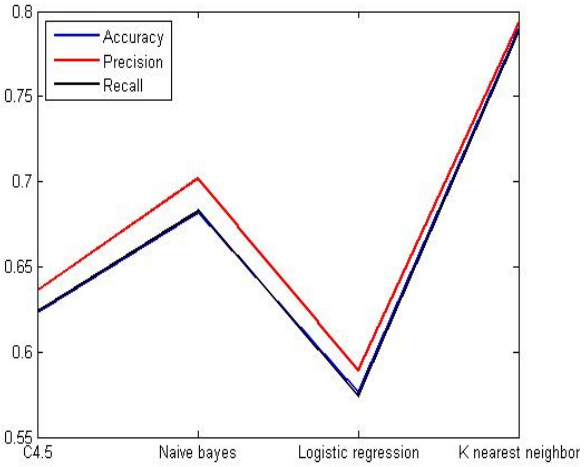

Figure 1. Accuracy chart on trace metal

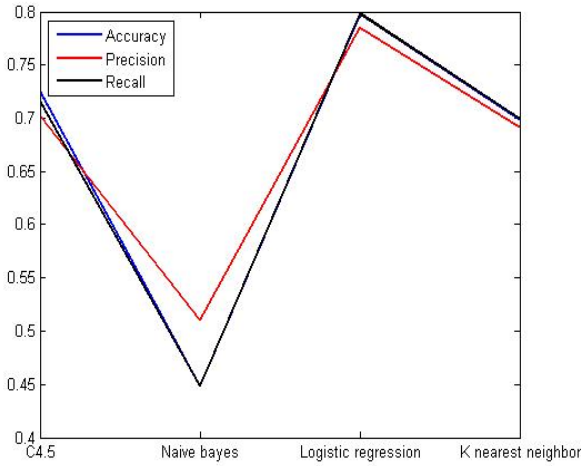

Figure 2. Accuracy chart on vehicle

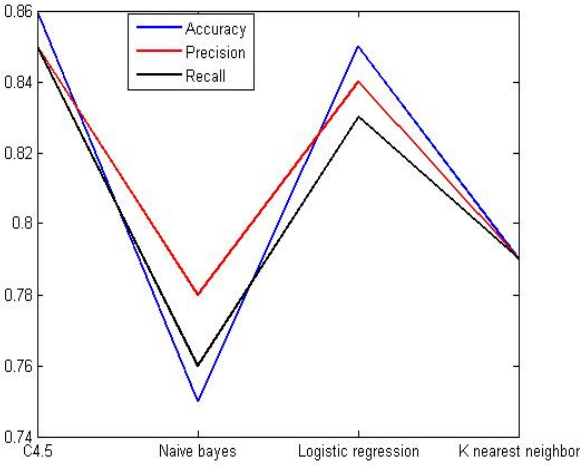

Figure 3. Accuracy chart on adults

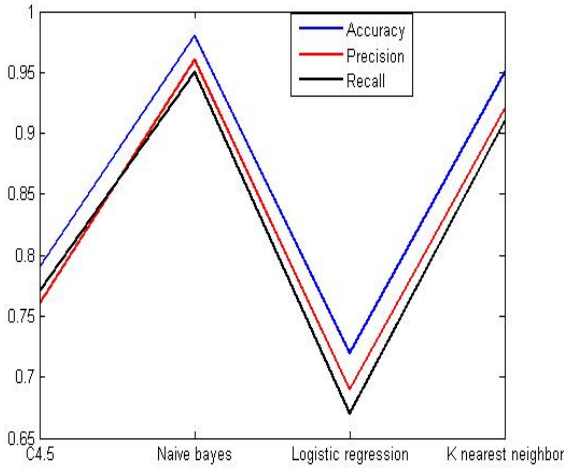

Figure 4. Accuracy chart on ionosonde

Conclusions

In this paper, we evaluate performance in terms of classification accuracy of C4.5, naive bayes, logistic regression, K nearest neighbor algorithms using various accuracy measures like precision, recall. Compare results of the first two datasets, K nearest neighbor and logistic regression have better accuracy. Trace metal dataset has less samples and attributes and vehicle dataset has less samples but more attributes. K nearest neighbor has mature theory, simple thinking, high accuracy and other advantages. Logistic regression has fast calculation. Compare results of the last two datasets, C4.5 and naive bayes have better accuracy, adults dataset has more samples and attributes and vehicle dataset has more samples but more missing value. In ionosonde dataset Naive bayes and K nearest neighbor is relatively simple and insensitive to missing value.

Acknowledgment

This research is financially supported by Henan Province Outstanding Youth on Science and Technology Innovation (No.164100510017); National 973 Program (No. 613237); National Natural Science Foundation of China (No. 61501407), respectively.

References

[1]Garcia S, Lu J, Herrera F. Data preprocessing in data mining[M]. Switzerland: Springer, 2015.

[2]Han J, Kambr M. Data Mining: Concepts and Techniques. Morgan Kaufman[M]. 2011.1-3.

[3]Juan H. Application of C4.5 algorithm in graduate enrollment[J]. 2015.

[4]Sudhakar A, Meleet M. A System for Extraction of Semantic Biomedical Relations Using Multi-nominal Naive Bayes Algorithm[J]. International Journal, 2014, 2(3).

[5]Ganser M, Dhar S, Kurup U. Patient Identification for Telehealth Programs[C]. International Conference on Machine Learning and Applications. 2015.

[6]Pu Y, Peng J, Huang L. An efficient KNN algorithm implemented on FPGA based heterogeneous computing system using OpenCL[C]. Field-Programmable Custom Computing Machines, 2015 IEEE 23rd Annual International Symposium, 2015: 167-170.

[7]http://archive.ics.uci.edu/ml/datasets.html

[8]http://www.chinare.org.cn/difDetailPublic/?id=3348