ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 168

International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)

Gaussian Naive Bayesian Data Classification Model Based on Clustering Algorithm

Zeng-jun BI*, Yao-quan HAN, Cai-quan HUANG and Min WANG

Air Force Early Warning Academy, Wuhan, China

*Corresponding author

Keywords: Clustering algorithm, Naive bayesian algorithm, Classification model.

Abstract. A gaussian naive bayesian data classification model based on clustering algorithm was proposed for fast recognition and classification of unknown continuous data containing a large number of non-priori knowledge. Firstly, the unknown data were extracted from the representative samples according to the information entropy measure for clustering to generate class labels. Then, the mapping relationship between data and class labels was established by using the gaussian naive bayes algorithm, and the classification model was obtained through training. Simulation results show that this unsupervised analysis process has a good classification effect on new data.

Introduction

Classification is an important part of data mining. By learning training data, the mapping relationship between training data and predefined classes can be established[1]. In order to make the traditional classification algorithm classify data well without predetermined classification for learning semi-supervised or even unsupervised methods are used to improve the classification algorithm[2]. Literature [3] uses semi-supervised naive bayes classification algorithm to establish initial classification for a small number of data sets with class labels, and continuously updates the data with high classification accuracy to the training set when predicting and classifying the data without labels, so as to realize semi-supervised learning of data classification. However, this algorithm fails to fundamentally realize the unsupervised generation of class labels of data to be classified, and prior knowledge still plays a crucial role in the training of classification algorithm. Clustering is an unsupervised process in which the most similar objects are divided into a class based on the objects found in the data and their relationships[4,5]; Literature [6] applies unsupervised clustering to text clustering and constructs an automatic text classification model based on vector space model. However, the model is not suitable for the classification of continuous variables.

Therefore, this paper combines the clustering algorithm with the gaussian naive bayes classification algorithm, and proposes an unsupervised classification model suitable for continuous variable data. In this method, small representative samples are extracted from large samples by information entropy theory, and prediction classes of observation data are generated by clustering algorithm as predefined target classes of classification algorithm, so that data are classified and prediction models are established without prior knowledge. Simulation results show that this model is efficient in classifying and processing new data, and only a small part of sample extraction is needed to train the classification model of the whole data, which greatly saves computing resources and time.

Selection of Clustering Algorithm

Classical clustering algorithm can be divided into hierarchical clustering algorithm, divide-based clustering algorithm and density-based clustering algorithm. The corresponding representative classical algorithms are k-means, condensed hierarchical clustering algorithm and DBSCAN.

Clustering performance measurement measures the performance of clustering algorithms under different environments according to the accuracy, consistency and other indicators of various clustering algorithms for sample division. ARI index is used to measure the consistency between the data label calculated by the clustering algorithm and the original label. The expression is:

$$ARI = \frac{RI\text{-}E[RI]}{\max RI\text{-}E[RI]} \quad (1)$$

Where, $RI = (a+b)/C_n^2$, $E[RI]$ is the variance of RI, and a and b respectively represent the number of data pairs that are still in the same class before and after clustering and the number of data pairs that are not in the same class. ARI index reflects the consistency between class tags obtained by clustering algorithm and self-carried class tags in data. Higher ARI value can reflect more accurate clustering performance.

ARI index was used to measure the clustering effect of k-means, hierarchical clustering and DBSCAN clustering algorithm on a sample size of 50-100. Generate three two-dimensional random point groups with [1.5,1.5], [2,2] and [3,3] as the center and subject to gaussian distribution, with sample size ranging from 50 to 100. The relation between ARI index and sample size of the three clustering algorithms is shown in Fig. 1.

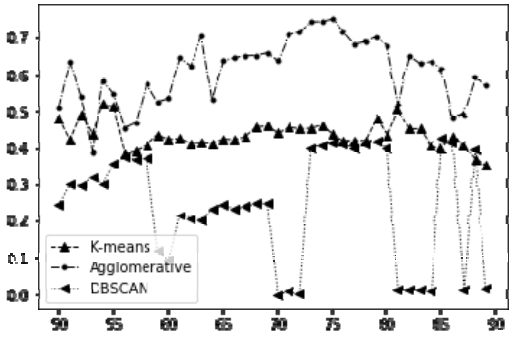


Figure 1. Relation between ARI index and sample size

As can be seen from figure 1, the ARI index of the hierarchical clustering algorithm between the sample size of 65-80 is the highest and relatively stable, which proves that the hierarchical clustering algorithm has the advantage of high accuracy in the small sample environment. Therefore, this paper will use this feature to take the hierarchical clustering algorithm as the clustering algorithm to generate the class tags of sample data.

Gaussian Naive Bayes Model Combined with Clustering Algorithm

In this paper, a data classification model combining the clustering algorithm and the gaussian bayesian classification algorithm is proposed. On the premise of no prior knowledge, a classification model is generated to assign an appropriate class tag Y to the sample DATA, and DATA of the same type, NEW_DATA_i, is classified according to the unified standard. The model building process is shown in Fig. 2.

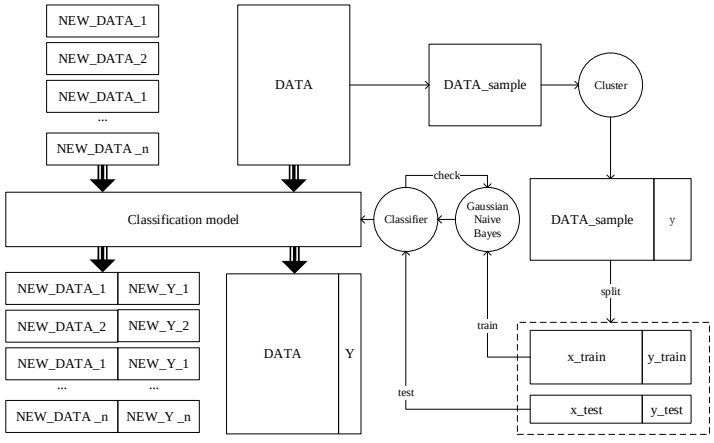


Figure 2. Model building process

The DATA_sample was extracted from large-scale DATA according to the information entropy as the learning DATA set of the model, and the DATA_sample was input into the clustering algorithm

to obtain the class tag y of the DATA_ sample. It can be considered that the correspondence between DATA_sample and class tag y can represent the correspondence between DATA and y.

DATA_sample was randomly split into training set (x_train) and test set (x_test) according to a certain ratio. Correspondingly, y was split into y_train and y_test.The x_train and y_train were input into the Gaussian Bayes algorithm for Classifier training. Test the classifier with x_test and y_test and check the classifier's accuracy. The calculation criterion of accuracy is the percentage of the number of samples that the predicted class label of the classifier on x_test is consistent with y_test.

The classifier is encapsulated to generate a classification model that masters the distribution and Classification rules of DATA. After the model is deployed, the corresponding class tag can be obtained according to the consistent principle for all sample data DATA or DATA with the same distribution NEW_DATA input into the model.

Experimental Simulation

This experiment was conducted on a 64-bit Windows 10 operating system with 8GB of computer memory. The algorithm uses Python language to compile and run on Jypyter software.

2000 sample points subject to gaussian distribution containing four classes were generated, and each point carried the original class label labels_true as the basis for model performance measurement. The sample distribution is shown in Fig. 3.

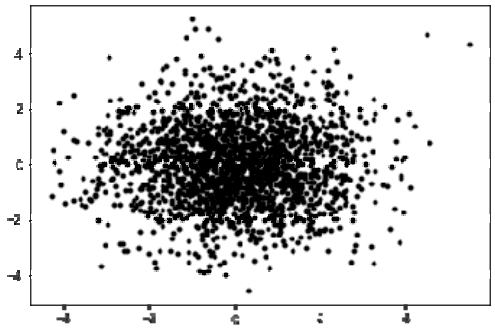


Figure 3. Sample distribution diagram

The entropy of information and the accuracy of corresponding classification algorithm under training sets of different scales are calculated as shown in Fig. 4.

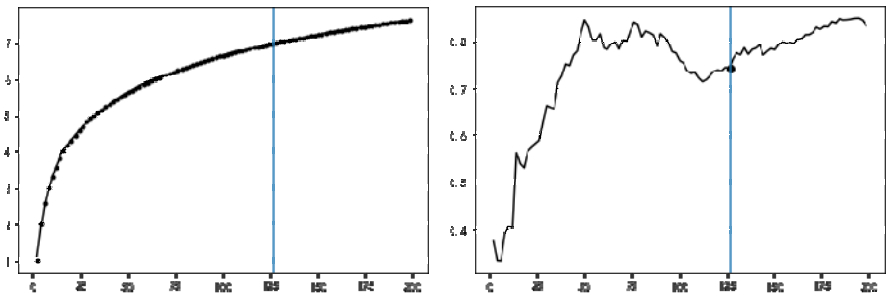


Figure 4. Information entropy and accuracy of corresponding classification algorithm under training sets of different scales

By calculation, the second derivative of the information entropy curve is zero when the training set size is 126, and the accuracy of the classification algorithm is 0.74.

126 sample points were extracted from 2000 sample data as the training set of the model. After clustering, 126 sample points were assigned with corresponding class labels. The distribution of training sets before and after clustering is shown in Fig. 5, where the sample points of different classes are expressed in different shapes after clustering.

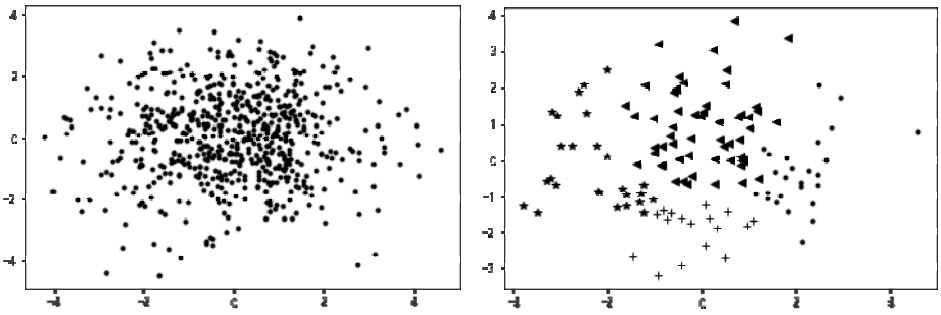
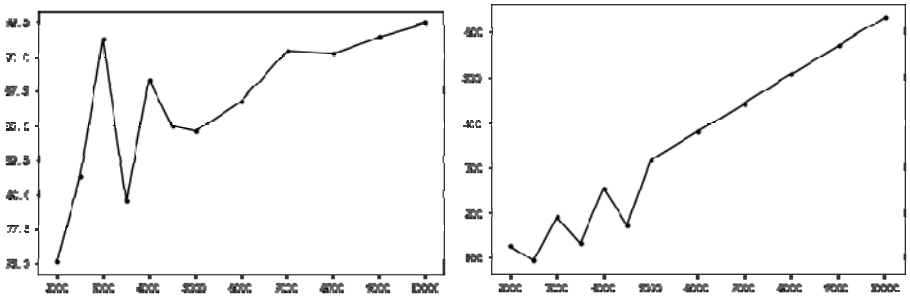


Figure 5. Training set distribution before and after clustering

The training set with the class tag after clustering is input into the naive bayes new algorithm for training, and the accuracy of the model obtained by the test set (x_test) is 0.75.

Further calculation shows that when the overall sample size increases, the advantages of this method in processing large sample data are more obvious, as shown in Fig. 6.



(a) (b)

Figure 6. Relationship between sample size and model accuracy and information entropy

As can be seen from Fig. 6(a), with the expansion of sample size, the accuracy of the model is constantly improved and in a high state. As can be seen from Fig. 6(b), information entropy, as a measure of training set size selection, enables samples of different sizes to reasonably select training sets reflecting the law of sample distribution, which not only ensures the accuracy of classification algorithm, but also controls the overall small size of training set relative to samples.

Conclusion

In order to realize unsupervised classification of unknown data, this paper proposes a continuous data analysis model combining clustering algorithm and naive bayesian classification algorithm. Using the information entropy theory drawn from the larger data sample smaller data set for the model study, based on hierarchical clustering algorithm under small data set is of high accuracy, and use of the advantages of hierarchical clustering algorithm for bayesian classification algorithm to generate the target class, makes the probability rule of bayes algorithm is effective to master data. Through the test of simulated data, it is proved that the method in this paper can only extract a small part of data in the sample to train the classification model of the overall data, which greatly saves the machine computing resources and model training time.

References

[1] Le mingming, research and application of data mining classification algorithm [D]. Chengdu university of electronic science and technology, 2017: 12-16.

[2] Kong yiqing, semi-supervised learning and its application research [D]. Wuxi, Jiangnan University, 2009: 33-39.

[3] Dong liyan, sui peng, sun peng, li yongli, a new naive bayesian algorithm based on semi-supervised learning [J]. Journal of Jilin University (engineering science edition), 2016, 46(3): 884-889.

[4] IEEE Translations on Power Systems, 2006, 21(2):933-940.

[5] Zhang bin, zhuang chijie, hu jun, et al. Power load curve integrated clustering algorithm combined with dimension reduction technology [J]. Chinese journal of electrical engineering. 2015. 35(15): 3741-3749.

[6] Zhu cuiling, research on text classification methods based on unsupervised clustering and naive bayes classification [D]. Jinan: Shandong University, 2005: 33-39.