Ali music trend prediction competition - the implementation of random forest algorithm

Zhenni Lu

1 CART

CART is composed of two parts: classification tree and regression tree. Classification tree is used for data analysis of discrete variables, and regression tree is used for data analysis of continuous variables. CART uses the Gini index as a split standard to select the attributes. When establish the CART tree, the selection of each split attribute is based on the degree to which the sample data is divided under different predictions.

2 Bagging

Breiman proposed the bagging algorithm in 1996. This algorithm generates the different component classifiers by operating the training sample set. The basis of Bagging algorithm is Bootstrap Sampling, that is, the training sample set T_set is randomly extracted from the original sample set S, and the number of samples in T_set is the same as S. The main idea of Bagging algorithm is: given a meta learning algorithm and an original sample set S, let the learning algorithm train for multi rounds, during each round the T_set training set was obtained from the original sample set S through the self-help sampling, such an initial sample may appear many times or not in the training in a round of training.

Assuming that N is the number of samples in the original sample set S, when the S is sampled using the Bootstrap method, the probability of not being extracted of each sample in S is $(1-1/N)^N$. If N is very large, $(1-1/N)^N \approx 0.368$, which shows that in the original S nearly 37% sample will not appear in the T_set. By constructing different training sets, the differences between the classifiers increase, and the Bagging method improves the generalization ability of the combined classifier. Through the T rounds training, we get a classifier sequence$\{f_1, f_2, \ldots, f_T\}$, then they form a multi classifier system, and the final classification results of the system are obtained by simple majority voting or simple average method.

3 Random forest

3.1 Introduction

Random forest is a new combinatorial classifier algorithm proposed by Breiman in 2001. A classification and regression tree (CART) is used as a meta classifier, and a training sample set is made by using Bagging method. When constructing a single tree, the attributes are randomly selected to split the internal nodes. The combination of Bagging algorithm and CART algorithm,

plus random selection features for attribute splitting, enables RF to tolerate noise better, so that it has better classification performance.

3.2 Definition

The Random forest is a collection of tree classifiers $\{h(x,\theta_k),k=1,\ldots\}$, the meta classifier $h(x,\theta_k)$ is a classification and regression tree constructed by CART algorithm without pruning. x is input vector, $\{\theta_k\}$ is an independent and identically distributed random vector, and the random vector θ_k determines the growth process of a single tree.

$\{h(x,\theta_k),k=1,\ldots\}$ in the traditional CART algorithm, each internal node is a subset of the original data set, the root node contains all the original data; in each of the internal nodes, find the best split way to split from all the attributes; then the subsequent nodes were divided until a leaf node; finally through the pruning minimum test error. Unlike the CART algorithm, the growth of a single tree in a random forest can be summarized as follows:

(1) use Bagging to form different training sets: assuming that the number of samples in the original training set is N, randomly select N samples from the original training set to form a new training set, so as to generate a classification tree.

(2) randomly selected features to split the classification and regression tree internal node: there are M features, specify a positive integer $m<<M$; in each internal node, from the M characteristic randomly select m features as candidate features, select the best split mode on these m features to split the nodes. During the growth of the whole forest, the value of M remains unchanged.

(3) every tree is free to grow and does not cut pruning.

The combinatorial method of random forest output has a simple majority voting method (for classification), and the average of the output results of a single tree (for regression).

3.3 Random forest algorithm flow

Algorithm: Random forest

Input: sample set $D=\{(x_1,y_1),(x_2,y_2),...(x_N,y_N)\}$, in which $x_i=(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(M)})^T$ is input instance (eigenvector). M is the number of features. m is number of input features which is used to determine the decision results of a node on a decision tree, m should be far less than M.

Output: Final classifier/regressor f(x)

1. From N training cases (samples), a training set is formed by random sampling(i.e. bootstrap sampling) N times in the way of return sampling, and the error is evaluated by the use cases (samples) that are not drawn.

2. For each node, m features are randomly selected, and the decision of each node on the decision

tree is determined based on these features. According to these m features, the best mode of splitting is calculated.

3. Each tree will grow intact without pruning, which is likely to be used after a normal tree classifier is built.

4. If it is classification algorithm prediction, the output combination method of the random forest is a simple majority voting method. If it is regression algorithm prediction, the output combination method of the random forest is the average of the output results of a single tree.

4 Experiment

4.1 Data

The data set consists of data from Ali music artist data, and the historical records of user behavior within 6 months (20150301-20150830) associated with these artists.

Table. 1: user-action table(mars_tianchi_user_actions)

Column name	Type	Description
User_id	string	User unique identifier
Song_id	string	Song unique identifier
Gmt_create	string	Playback time of users
Action_type	string	Action type:1,play; 2,download; 3,collect
Ds	string	Record date

Table. 2: song-artist table(mars_tianchi_songs)

Column name	Type	Description
Song_id	string	Song unique identifier
Artist_id	string	The artist of the song
Publish_time	string	The publish time of the song
song_init_plays	string	The initial play times of the song
Language	string	The language of the song:1,2,3
Gender	string	The type of the artist:1,men;2,women;3,band

The result is the prediction of the play data of artist in the next 2 months, that is, 60 days (20150701-20150830).

Table. 2: prediction results table

Column name	Type	Description
Artist_id	string	The artist of the song

plays	string	Playback data of the artist on the day
Ds	string	Date

4.1.1 Data processing

We use pandas to process and merge raw data into every row including the following attributes: user_id,song_id,Ds,play_num,download_num,collect_num,artist_id,publish_time,song_init_plays, Language,gender. At the same time we processed the Ds column, adding the is_weekend, is_festival two column attributes.

4.1.2 Feature selection

For the perspective of user-song, we use is_weekend, is_festival, download_num, collect_num, artist_id, publish_time, song_init_plays, Language and Gender as features.

4.2 Setup

The range of number of random forest neutron tree is set to 1-200, and select the number of subtrees with minimum mean square variance. According to the results of the experiment, the number of subtrees is set to 35, the max depth of tree is set to 12, and the max features set to 4 can reach the best result.

4.3 Evaluation

The actual playback number of artist j on day k is $T_{j,k}$, the artist set is W. The predicted playback number of artist j on day k is $S_{j,k}$, the normalized variance of prediction of the playback of artist j and the actual value is:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{k=1}^{N}((S_{j,k} - T_{j,k})/(T_{j,k}))^2}$$

The weight of the artist j is calculated according to the square root of the entertainer's playback:

$$\emptyset_j = \sqrt{\sum_{k=1}^{N} T_{j,k}}$$

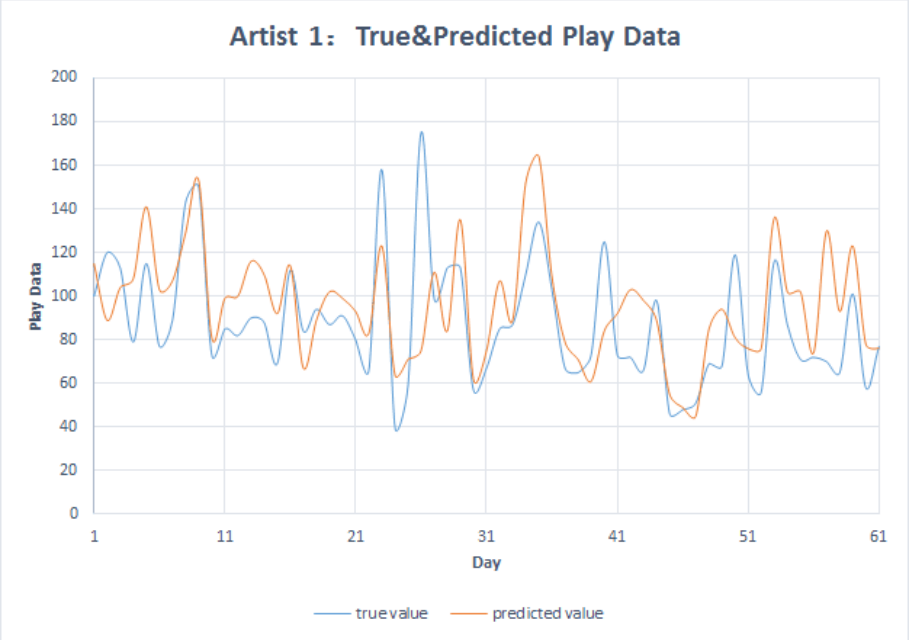
The evaluation index is F, and the greater the F value, the better the result:

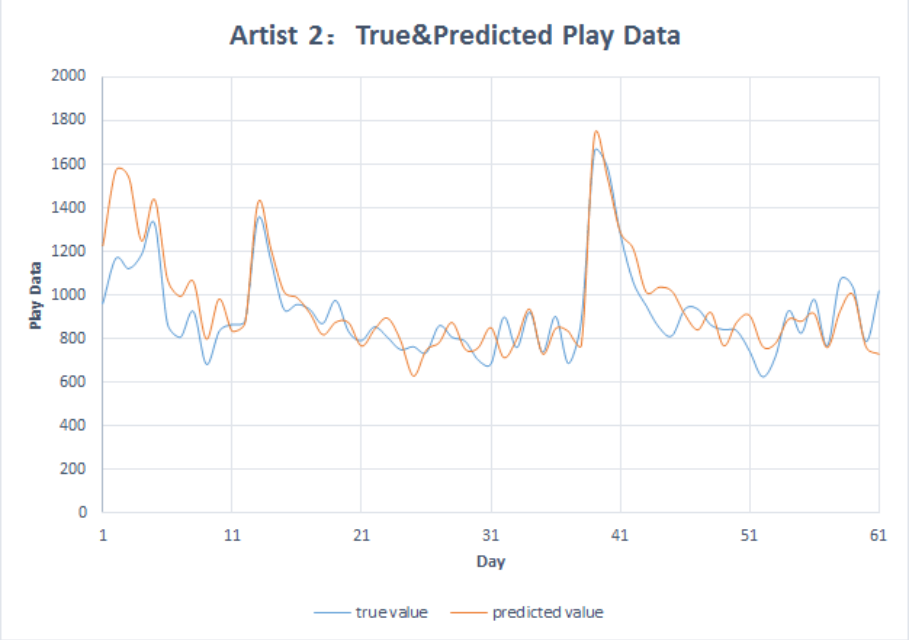
$$\mathrm{F} = \sum_{j \in W} (1 - \sigma_j) * \emptyset_j$$

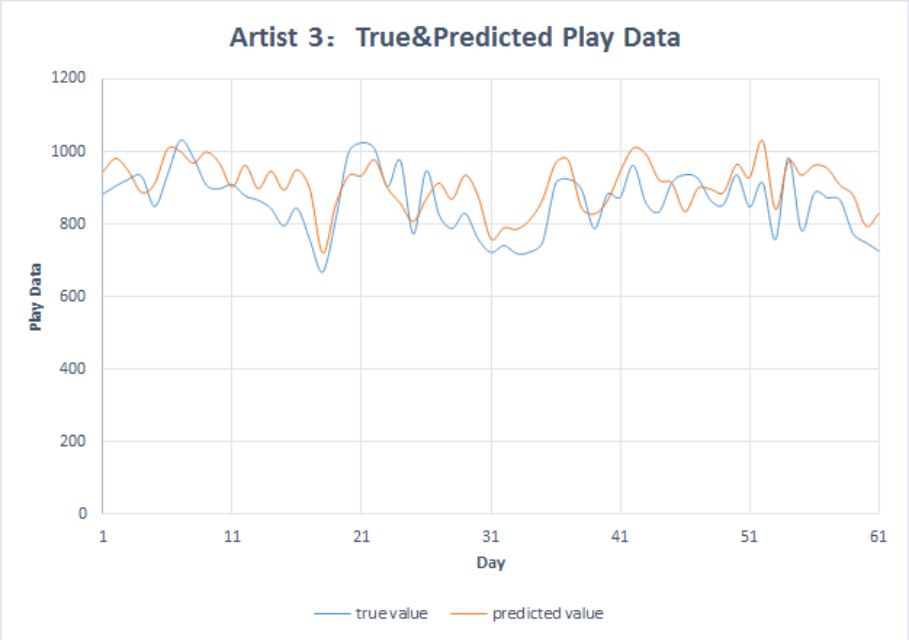
4.4 Result

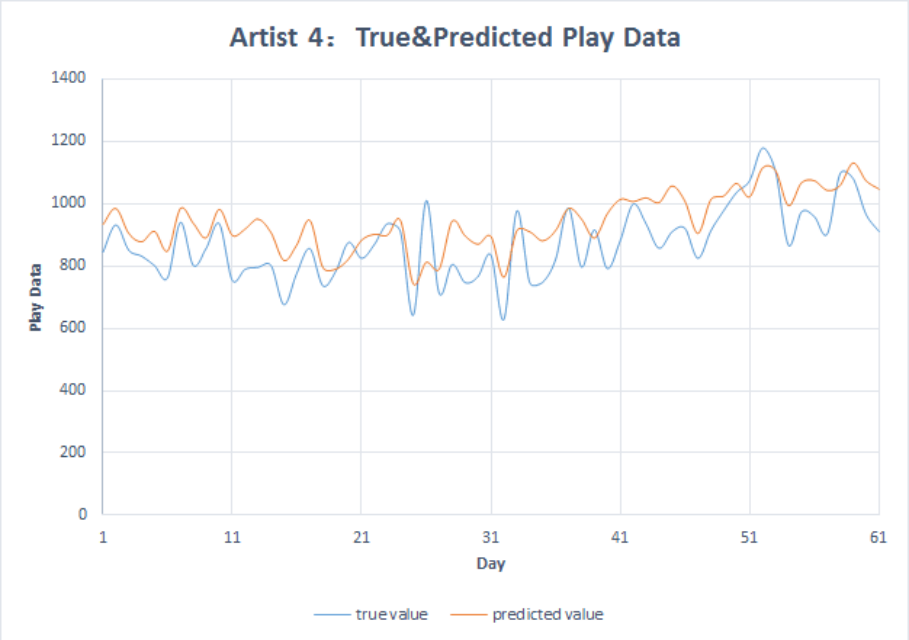
According to the experiment, the number of subtrees is set to 35, max depth of tree is set to 12, max features is set to 4 can reach the best result is best, and it is 6931.58625645. The following picture

show contrast between the real and predicted playback data of a few artists.






Artist 3：True&Predicted Play Data
1200
1000
800
600
400
200
0
Play Data
1
11
21
31
41
51
61
Day
true value
predicted value


Artist 4：True&Predicted Play Data
1400
1200
1000
800
600
400
200
0
Play Data
1
11
21
31
41
51
61
Day
true value
predicted value