Time series classification based on arima and adaboost

Jinghui Wang[1,*], and *Shugang* Tang[2]

[1]Key Laboratory of Intelligence Computing and Novel Software Technology
[2]Tianjin University of Technology, TianJin, China, 300384

Keywords: Time series classification, ARIMA, AdaBoost.

Abstract. In this paper, a novel time series classification approach, which using auto regressive integrated moving average model (ARIMA) features and Adaptive Boosting (AdaBoost) classifications. ARIMA is particularly suitable for distinguishing time series signal and Adaboost is suitable for features classification. The simulation results have shown that the algorithm is feasible. And this method is more accurate than many existing method in multiple time series problems.

1 Introduction

Time-series data arise in many fields including finance, signal processing, speech recognition and medicine. A standard approach of time-series problems usually requires feed time series features into a machine learning algorithm and get goal result [1]. Engineering of features generally requires some domain knowledge of the discipline where the data has originated from. For example, if one is dealing with signals (i.e. classification of EEG signals), then possible features would involve power spectra at various frequency bands and several other specialized statistical properties.

Time series classification(TSC) is a difficult problem, and the time series featrues are more unstable than image feature, so it's difficult to extract key features from them. And feature extraction of time series affects the final classification effect, so the current accuracy of the results is still very low.

At present, researchers often divide time series classification into three categories: classification by time point[2], classification by shapes and classification by change. The similarity in time series can be used distance, such as Euclidean distance. The spatial similarity can be used with DTW. The similarity of the data generation process can be used with the distance of probability, such as GMM, and ARMA's mixture.

In statistics and machine learning field, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms.

In this article, we will discuss auto regressive integrated moving average model(ARIMA)

* Corresponding author: Csr_dsp@sina.com

[3] and AdaBoost [4] methods to classify time-series data, The ARIMA process acts as a feature exatrcation and AdaBoost acts as a classifier. We used UCI data set in the simulation test. Those data include the classic Human Activity Recognition (HAR) and Food dataset. The dataset is the raw time-series data. We compared the performance of ARIMA and AdaBoost classification with convolutional and recurrent neural networks. The simulation test show that the times series classification based on ARIMA and Adaboost can work, and it is particularly suitable for stationary time series signals, and the result is relatively very well.

2 Time series classification model

Time series classification problems are differentiated from traditional classification problems because the attributes are ordered and long. Whether the ordering is by time or not, it is in fact related with each other. The important characteristic is that there important features dependent on the ordering.

Classification, which is the task of assigning objects to one of several predefined categories, which is a pervasive problem that encompasses many diverse applications.

Definition (Classification).Classification is the task of learning a target function that maps each attribute set x to one of the predefined class labels y. The target function is knowns informally as a classification model.

A case is a pair{x,y} with m observations $x_1, \dots, x_m$ (the time series) and discrete class variable y, which is c possible values. n represents the number of sample points.

$$\mathrm{T} =< \mathrm{X}, \mathrm{y} >=< \left(x_{1,} y_{1,}\right), \dots, \left(x_{n,} y_{n,}\right) > \tag{1}$$

A classifier is a function or mapping from the space of possible inputs to a probability distribution over the class variable values, and time series classification algorithms involve some processing and filtering of the original time series. As with general classification, our goal is to classify time series.

3. Time series classification model arima and adaboost

3.1 Auto regressive integrated moving average model

ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and provides a simple powerful method for making skillful time series forecasts[5].This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

q: The size of the moving average window, also called the order of moving average.

An auto regressive (AR(p)) component is referring to the use of past values in the regression equation for the series Y. The auto-regressive parameter p specifies the number of lags used in the model. For example, AR(2) or, equivalently, ARIMA(2,0,0), is represented as

$$Y_t = \text{c} + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + e_t \tag{2}$$

where φ1, φ2 are parameters for the model.

The d represents the degree of differencing in the integrated (I(d)) component. Differencing a series involves simply subtracting its current and previous values d times. Often, differencing is used to stabilize the series when the stationarity assumption is not met, which we will discuss below.

A moving average (MA(q)) component represents the error of the model as a combination of previous error terms et. The order q determines the number of terms to include in the model

$$Y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t \tag{3}$$

where y_d is Y differenced d times and c is a constant.

ARIMA models can be also specified through a seasonal structure. In this case, the model is specified by order parameters: (p, d, q)

3.2 AdaBoost algorithm

AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. The final equation for classification can be represented as[6]

$$F(x) = sign\left(\sum_{m=1}^{M} \theta_m f_m(x)\right) \tag{4}$$

where f_m stands for the m_th weak classifier and θ_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers. The whole procedure of the AdaBoost algorithm can be summarized as follow.

Given: $(x_1, y_1) \ldots, (x_m, y_m)$ where $x_i \in X, y_i \in \{-1, +1\}$.

Initialize: $D_1(i) = \frac{1}{m}$ for $i = 1, \ldots, m$.

For $t = 1, \ldots, T$:

(1) Train weak learner using distribution D_t

(2) Get weak hypothesis $h_t: X \to \{-1, +1\}$

(3) Aim:select h_t with low weighted error:

$$\varepsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$$

(4) Choose $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

(5) Update,for $i = 1, \ldots, m$:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$\mathrm{H(x)} = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$$

4 Time series classification based on arima and adaboost

4.1 Algorithm flow

The algorithm flow as follows:
(1) Choose the values of p,d and q.
(2) Using the ARIMA method to calculate the coefficient (p, d, q).
(3) Use coefficient (p,d,q) as train dataset and Adaboost as classifier to train time series classification, we can get AdaBoost train model.
(4) For the new time series, we can calculate the ARIMA coefficients (p, d, q), and using the trained AdaBoost model classification.

4.2 Alogrithm analysis

The algorithm based on ARIMA and AdaBoost. The inherent nature of ARIMA and adaboost determines some of the issues that arise from using the algorithm, such as:

(1) If the series has positive autocorrelations out to a high number of lags, then it probably needs a higher order of differencing;(2) The optimal order of differencing is often the order of differencing at which the standard deviation is lowest; (3) A model with no orders of differencing assumes that the original series is stationary (among other things, mean-reverting) (4) How to decide the max depth of the decision tree?(5) How to decide the min samples leaf of the decision tree?

In practice, different values have different effects, and we can debug parameters to determine the values used. Also,we can solve this problem in different ways.

5. Simulation and synthetic signals

In order to examine the method, we use the UCR datasets [7].

5.1 Two categories problem

The two categories problem example used data is the classic Food spectrographs dataset, which from the UCR repository. The dataset are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. The coffee dataset is a two-class problem to distinguish between Robusta and Aribica coffee beans. Further information can be found in the original paper [8]. The data were first used in the time series classification literature by Bagnall et al. The time series diagram is shown in Figure 1.

The dataset characteristics are as follows:

Train size: 28; Test size: 28; Missing value: No; Number of classses: 2; Time series length 286.

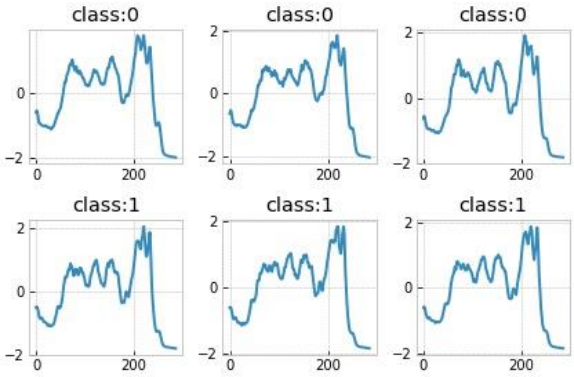

Fig. 1. Coffee Dataset and Class graphic.

From the above graph, we can find the time series waveforms of class 1 and class 2 are different, but it is difficult to clearly explain their differences.

From the Table I, we can see that the error rate is still relatively high. The main problem is that more class 1 is misclassified to class 2. The main problem is the problem of dtw calculation.

AB(1): ARIMA(5,1,0) ;AdaBoost: n_estimators=20;learning_rate=1;algorithm='SAMME.R'

AB(2): ARIMA(10,5,0) ;AdaBoost: n_estimators=50,;learning_rate=1;algorithm='SAMME.R'

Classification Accuracy(Two categories)

Table 1. Classification Accuracy(Two categories).

Method	Accurate ratio
MCNN	0.964([7])
BOSSVS	0.964([7])
DTW	0.811
AB(1)	0.941
AB(2)	0.952

5.2 Multiple classification problem

The multiple classification example used data is the Haptics dataset, which from the UCR repository. The dataset are taken from 5 people entering their passgraph (a code to access a system protected by a graphical authentication system) on a touchscreen. The data is the x-axis movement only.

The dataset characteristics are as follows:

Train size: 155; Test size: 308; Number of classses: 5; Time series length: 1092.

The graphs of these signals are very similar, and the data frame is longer than example I. So the multi-classification problems will be more difficult than the two-categories problem.

AB(3): ARIMA(5,1,0) AdaBoost: n_estimators=20, learning_rate=1,algorithm='SAMME.R'

AB(4): ARIMA(10,5,0) AdaBoost: n_estimators=20, learning_rate=1,algorithm='SAMME.R'

Table 2. Classification Accuracy (Multiple categories).

Method	Accurate ratio
MCNN	0.47
BOSSVS	0.416
DTW	0.377
AB(3)	0.477
AB(4)	0.502

From the above classification accuracy table II, we can find the classification algorithm is better for some classes, and it's worse for others. Overall classification is better than random guessing, but there many problem in practical application.

6 Summary

In this paper, we have presented time series classifier based on ARIMA and AdaBoost. Simulation results have shown that the method is feasible. The method performs better than the benchmark neural network.

At the same time, there some problems exists, such as described above. In the following work, we need to solve the problems including: (1) The ARIMA calculation is still slow, we try to find a faster method than above ARIMA. A faster algorithm can save computing resources and computing time; (2) We try to test whether the ARIMA coefficient can effectively represent the characteristics of time series, especially those of time series with very small differences. (3) How to improve classification accuracy of one-versus-rest? (4) We need to find relationship between the ARIMA coefficients and AdaBoost parameter to the classification accuracy.

This work is supported by National Natural Science Foundation of China (No.61001174),

References

1. P. Marteau.TimeWarp Edit Distance with Stiffness Adjustment for Time Series Matching.IEEE Transactions on Pattern Analysis and Machine Intelligence 31,2,306-318. 2008.
2. Hyndman, Rob J; Athanasopoulos, George. 8.9 Seasonal ARIMA models. Forecasting: principles and practice. oTexts. Retrieved 19 May 2015.
3. Freund, Yoav; Schapire, Robert E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. 1995.
4. D.Goldin and P.Kanellakis,On similarity queries for time-seriesdata: Constraint specification and implementation, 1995, pp. 137–153.
5. Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3), 263-286, 2001
6. W.B.H. Ali, R. Nock, M. Barlaud, Boosting Stochastic Newton with Entropy Constraint for Large-Scale Image Classification, International Conference on Pattern Recognition Stockholm Sweden August, pp. 232-237, 2014.
7. Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi , Chotirat Ann Ratanamahatana, Yanping Chen, Bing Hu, Nurjahan Begum, Anthony Bagnall , Abdullah Mueen and Gustavo Batista (2018). The

UCR Time Series Classification Archive. URL https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

8. Wang, Zhiguang & Yan, Weizhong & Oates, Tim. (2017). Time series classification from scratch with deep neural networks: A strong baseline. 1578-1585. 10.1109/IJCNN.2017.7966039.
9. Bagnall, Anthony, et al. Transformation based ensembles for time series classification. Proceedings of the 2012 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2012.