Classification of Multifractal Time Series by Decision Tree Methods

Bulakh Vitalii[1], Lyudmyla Kirichenko[0000-0002-2426-0777], Tamara Radivilova[0000-0001-5975-0269]

Kharkiv National University of Radioelectronics, Kharkiv, 61166, Ukraine
bulakhvitalii@gmail.com
lyudmyla.kirichenko@nure.ua
tamara.radivilova@gmail.com

Abstract. The article considers classification task of model fractal time series by the methods of machine learning. To classify the series, it is proposed to use the meta algorithms based on decision trees. To modeling the fractal time series, binomial stochastic cascade processes are used. Classification of time series by the ensembles of decision trees models is carried out. The analysis indicates that the best results are obtained by the methods of bagging and random forest which use regression trees.

Keywords: multifractal time series, binomial stochastic cascade, classification of time series, Random Forest, Bagging

1 Introduction

Many complex systems have a fractal structure and their dynamics is represented by time series that have fractal (self-similar) properties. Fractal time series take place in technical, physical, biological and information systems. The results of time series fractal analysis are widely used in various fields, in particular, in recognition and classification.

The tasks of classification of fractal realizations arise when medical diagnosis clarification by ECG and EEG [1], when DDoS-attacks are detected by the incoming traffic [2], when critical situations in financial markets are forecasted by economic indicators series [3], etc. Most often, such tasks are solved by estimating and analyzing fractal characteristics. However, in recent years, there has been a growing interest in machine learning methods to analyze and classify fractal series [4-6].

Key to correctly solving the classification problem is the choice of the classification method. To answer the question which method is best for analyzing multifractal properties, we present the results of study in which the classification of model realizations possessing fractal properties was carried out. The aim of the work is a comparative analysis of the classification of multifractal stochastic time series performed by methods based on decision trees.

2 Multifractal Time Series

The self-similarity of random process is to preserve distribution law when changing the time scale. A stochastic process $X(t)$ is self-similar with a parameter H if the process $a^{-H}X(at)$ is described by the same finite-dimensional distribution laws as $X(t)$. The parameter H, $0 < H < 1$, called the Hurst exponent, represents the measure of self-similarity and the long-term dependence. Multifractal stochastic processes are inhomogeneous self-similar ones and have more flexible scaling relations. The multifractal properties of process are defined the scaling exponent $\tau(q)$. In the general case $\tau(q)$ is a nonlinear function for which the value $(\tau+1)/2$ coincides with the value of Hurst exponent H. For monofractal process $\tau(q)$ is linear. [7].

One of the frequently used models of the multifractal process is the conservative stochastic binomial multiplicative cascade. To its construction, an iterative procedure based on two main parts is used. The first represents geometric detailing by iterative partitioning of intervals, and the second guarantees randomness of the weighting coefficients. For each iteration n, $n \geq 1$, we have a time series (cascade) with multifractal properties.

The fractal characteristics of stochastic multiplicative cascade obtained using beta distribution random variable $Beta(\alpha, \beta)$ are completely determined by the parameters $\alpha, \beta > 0$ [8]. The change of value α of the symmetric distribution $Beta(\alpha, \alpha)$, allows to generate of multiplicative cascades with specified multifractal properties and Hurst parameter in the range $0.5 < H < 1$.

3 Classification Methods

The decision tree method is effective method of classification. It is applicable to solving classification problems arising in various fields. It consists in the process of partition the original data set into groups, until homogeneous subsets are obtained. The set of rules that give such a partition allows then make a conclusion for new data. However the decision tree models are unstable: a slightest change in the training set can bring to the essential changes in the tree structure. In this case, it is expedient to use ensembles of elementary classifiers. The components of the ensemble can be the same type or different.

The bagging method [9] is based on the statistical method of bootstrap aggregating. Bagging is a classification technique where all the elementary classifiers are trained and operate independently of each other. The basis of the bagging method is the classification technology, called "perturbation and combination". Perturbation is understood as the introducing of some random changes in training data and the construction of several alternative models on the modified data. From a single training set several samples containing the same number of objects are extracted by sampling. To obtain the result of the work of the ensemble of models, the voting or averaging are usually used. The effectiveness of bagging is achieved due to the fact that the basic algo-

rithms, trained in different subsamples, are obtained quite different and their mistakes are mutually compensated in the voting process.

Random forest is also a method of bagging, but it has several features [10]: it uses an ensemble of only regression or classifying decision trees; in the sampling algorithm the random selection of features is also carried; the decision trees are built up to the full completion of the training objects and are not subjected to post pruning.

4 Research results

To build decision tree models, Python with libraries that implement machine learning methods was used [11]. Classification of time series obtained by generating stochastic binomial cascades with different multifractal properties was carried out. Each class was a set of model time series with the same Hurst exponent. Hurst exponent values were varied in the range of 0.5 to 1 in increments of 0.05. Thus, the training of models was carried out in 11 classes.

In the work, to determine the time series belonging to one of the 11 classes, the methods of bagging and random forest were used. In this case the objects were the cascade time series, and the features were the values of this series. In each of the methods, the ensembles of decision trees, both classifications and regressions, were involved. In the case of regression decision trees, the result of the classification is the probability of matching of multifractal cascade to given class. The models for each class were trained on five hundred examples of time series and were tested on fifty test ones.

The probabilities are calculated by the formula: $P_i = 1 - |m_i - C|$, where m_i is the regression result for the *i*-th example, *C* is theoretically known class number. If the condition $P_i \in [0.5;1]$ is met, the classification is considered as correct. If $P_i < 0.5$ and $m_i > C$ then the cluster number is overvalued, otherwise it is understated.

Table 1 presents the average probabilities of class determination depending on the length of time series and the method of classification. The results show that the use of regression trees gives significantly greater accuracy than classification trees. The random forest method showed better results than bagging. It should be noted that random forest correctly classifies cascades of different lengths, what allows it to be used to classify short time series.

Table 1. Average probability of class determination

Length of time series	Bagging classification	Bagging regression	Random forest classification	Random forest regression
512	0.646	0.788	0.806	0.85
1024	0.655	0.832	0.834	0.878
2048	0.676	0.882	0.842	0.916
4096	0.71	0.896	0.852	0.918
8192	0.748	0.9	0.866	0.922
16384	0.768	0.93	0.872	0.926

5 Conclusion

In this paper, the comparative analysis of the classification of model multifractal stochastic time series using meta-algorithms based on decision trees has been performed. Binomial multiplicative stochastic cascades were used as input time series.

Time series were divided into classes depending on their fractal properties. Random forest and bagging methods were used to classify the series. In each method, ensembles of decision trees, both of classification and regression, were involved.

From the research that has been carried it is possible to conclude that the classification of series by fractal properties using decision trees methods gives good results. The best results were obtained with the use of regression trees. In the classification of the series with a small length random forest method showed greater accuracy.

The obtained results can be used for practical applications related to the classification or clustering of real time series with fractal properties. In our future researches we intend to concentrate on the classification of real series using additional features such as fractal characteristics.

References

1. Alghawli, A., Kirichenko, L.: Multifractal Properties of Bioelectric Signals under Various Physiological States. Information Content & Processing International Journal 2(2), 138-163 (2015)
2. Kaur, G., Saxena, V., Gupta, J.: Detection of TCP targeted high bandwidth attacks using self-similarity. Journal of King Saud University - Computer and Information Sciences, 1-15 (2017).
3. Kristoufek, L.: Fractal Markets Hypothesis and the Global Financial Crisis: Scaling, Investment Horizons and Liquidity. https://arxiv.org/pdf/1203.4979.pdf (2012) last accessed 2018/03/26.
4. André, L., Coelho, V., Clodoaldo, A., Lima, M.: Assessing fractal dimension methods as feature extractors for EMG signal classification. Engineering Applications of Artificial Intelligence 36, 81–98 (2014).
5. Symeonidis, S.: Sentiment analysis via fractal dimension. In: Proceedings of the 6th Symposium on Future Directions in Information Access, 48-50 (2015).
6. Arjunan, S. P., Kumar, D. K., Naik, G. R.: A machine learning based method for classification of fractal features of forearm sEMG using Twin Support vector machines. In: Annual International Conference of the IEEE Engineering in Medicine and Biology, 4821-4824 (2010).
7. Riedi, R.H.: Multifractal processes, in Doukhan P., Oppenheim G., Taqqu M.S. (Eds.), Long Range Dependence: Theory and Applications: Birkhuser. 625–715 (2002).
8. Kirichenko, L., Radivilova, T., Kayali, E.: Modeling telecommunications traffic using the stochastic multifractal cascade process. Problems of Computer Intellectualization, 55–63 (2012).
9. Breiman, L.: Bagging predictors. Machine Learning. 24 (2), 123–140 (1996).
10. Breiman, L.: Random Forests. Machine Learning 45 (1), 5–32 (2001).
11. Cielen, D., Meysman, A., Ali, M.: Introducing Data Science: Big Data, Machine Learning, and more, using Python tools. Manning Publications (2016).