УДК 519.2:004.9

В.А. Булах, Л.О. Кириченко, Т.А. Радивилова

КЛАССИФИКАЦИЯ МУЛЬТИФРАКТАЛЬНЫХ СТОХАСТИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ МЕТА-АЛГОРИТМОВ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ

Аннотация. В статье проведен сравнительный анализ классификации фрактальных временных рядов с помощью мета-алгоритмов на основе деревьев решений. Для построения модельных фрактальных временных рядов были выбраны биномиальные стохастические каскадные процессы. Анализ показывает, что наилучшие результаты получены методами Случайного леса и Бэггинг, которые используют деревья регрессии.

Ключевые слова: мультифрактальные временные ряды, биномиальный стохастический каскад, классификация временных рядов, Случайный лес, Беггинг.

Введение и цель

Временные ряды являются основными данными для понимания динамики в реальных сложных системах разного типа. Для множества задач анализа временных рядов, в том числе для классификации, в настоящее время используются методы машинного обучения. Для разработки и тестирования новых методов важными являются математические модели, позволяющие генерировать модельные ряды с заданными свойствами для решения задач дата манинг [1-3].

Многие сложные системы обладают фрактальной структурой и их динамика представлена временными рядами, обладающими фрактальными свойствами. Результаты фрактального анализа временных рядов широко используются в технических, физических, биологических, информационных системах. Во многих случаях возникают задачи распознавания и классификации фрактальных рядов. Чаще всего это происходит путем оценивания и анализа фрактальных характеристик [4-5]. Однако в последние годы для анализа и классификации фрактальных рядов используются методы машинного обучения [6-8]. Целью представленной работы является сравнительный анализ классификации мультифрактальных стохастических временных рядов, выполняемых методами, основанными на деревьях решений.

Мультифрактальные временные ряды

Самоподобие случайных процессов заключается в сохранении статистических характеристик при изменении масштаба времени. Стохастический процесс $X(t)$ является самоподобным с параметром H, если процесс $a^{-H}X(at)$ описывается теми же конечномерными законами распределений, что и $X(t)$:

$$\mathrm{Law}\{a^{-H}X(a\,t)\} = \mathrm{Law}\{X(t)\}, \quad \forall a > 0,\ t > 0\ . \qquad (1)$$

Для мультифрактальных процессов рассматривается более общее соотношение

$$\mathrm{Law}\{X(a\,t)\} = \mathrm{Law}\{M(a)\cdot X(t)\}, \qquad (2)$$

где $M(a)$- независимая от $X(t)$ случайная функция. Мультифрактальные процессы имеют следующие скейлинговые закономерности для моментных характеристик:

$$\mathrm{M}\left[\left|X(t)\right|^{q}\right] \propto t^{\tau(q)+1}, \qquad (3)$$

где $\tau(q)$ - скейлинговая экспонента, в общем случае нелинейная функция, для которой значение $\tau + 1/q$ при $q = 2$ совпадает со значением показателя H.

Простыми моделями мультифрактального процесса с заданными свойствами являются биномиальные мультипликативные каскады [9,10]. При их построении первоначальный единичный отрезок делится на два равных интервала, которым приписываются весовые коэффициенты w_1 и $1-w_1$, где w_1 является значением некоторой заданной случайной величины W. На втором шаге добавляются два новых независимых случайных значения w_2 и w_3. Получится 4 интервала с весовыми коэффициентами w_1w_2, $w_1(1-w_2)$, $(1-w_1)w_3$ и $(1-w_1)(1-w_3)$. При $n \to \infty$ мы приходим к предельной мере, являющейся неоднородным фрактальным множеством.

Фрактальные характеристики биномиального стохастического мультипликативного каскада при использовании случайной величины бета-распределения задается его параметрами α и β. Для построения каскада с заданными фрактальными характеристиками необходимо решить уравнение:

$$\tau(q) = -\log_2 M\left[X^q\right] - 1 \tag{4}$$

где: $M\left[X^q\right] = \dfrac{B(\alpha+q,\beta)}{B(\alpha,\beta)}$ – начальные моменты случайной величины X порядка q, при этом $q > -\alpha$.

Данное уравнение неразрешимо в явном виде, поэтому для получения значений параметров α и β для существующей функции $\tau(q)$ необходимо решить условную оптимизационную задачу вида:

$$\alpha,\beta = \arg\min_{\alpha,\beta}\left(\tau(q) + \log_2 \frac{B(\alpha+q,\beta)}{B(\alpha,\beta)} + 1\right)^2 \to \min \tag{5}$$

с ограничениями: $\{\alpha > 0, \beta > 0, q > -\alpha\}$

Решение задачи (2) позволяет получить параметры бета-распределения, что дает возможность генерировать мультиприкативные каскады с параметром Херста в интервале от 0.5 до 1.

Методы исследования

Одним из самых эффективных методов для решения задач классификации, возникающих в самых разных областях, считается метод деревьев решений. Он состоит в том, чтобы осуществлять процесс деления исходных данных на группы, пока не будут получены однородные их подмножества. Совокупность правил, которые дают такое разбиение, позволяет затем делать вывод для новых данных.

Модели деревьев решений, неустойчивы: даже небольшое изменение в обучающем множестве может привезти к существенным изменениям в структуре дерева. В этом случае целесообразно использовать ансамбли моделей. Одним из первых и самых известных видов ансамблей является метод Бэггинг (Bagging). Бэггинг – технология классификации, где все элементарные классификаторы обучаются и работают независимо друг от друга. Идея заключается в том, что классификаторы не исправляют ошибки друг друга, а компенсируют их при голосовании.

В основе работы бэггинга лежит технология классификации, получившая название «возмущение и комбинирование» [11]. Под возмущением понимается внесение некоторых случайных изменений в обучающие данные и построение нескольких альтернативных моде-

 ISSN 1562-9945

лей на измененных данных с последующим комбинированием результата. Если ансамбль строится на основе моделей различных типов, то для каждого типа будет свой алгоритм обучения.

Для получения результата работы ансамбля моделей обычно используются следующие способы комбинирования: голосование или усреднение, которое может определяться как простое среднее значение выходов всех моделей. Эффективность бэггинга достигается благодаря тому, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при голосовании, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.

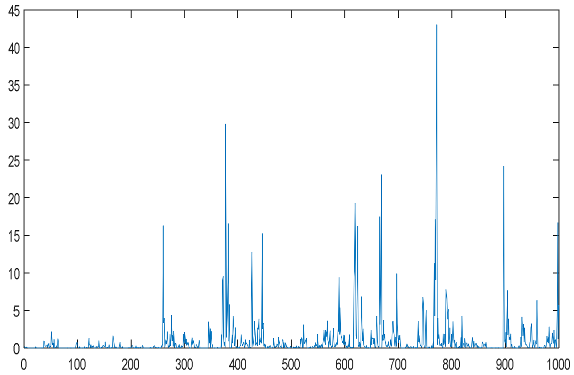
Случайный лес (Random Forest) также является методом Бэггинга, но в отличие от его основной версии имеет несколько особенностей: использует внутри себя ансамбль только регрессионных или классифицирующих деревьев решений; в алгоритме сэмплирования помимо выбора случайным образом обучающих объектов, также производится выбор случайным образом признаков; для каждой подвыборки дерево решений строится до полного исчерпания обучающих примеров и не подвергается процедуре отсечения ветвей [12].

Результаты исследования

Для построения моделей деревьев решений использовался язык Python с библиотеками, реализующими методы машинного обучения [13]. Классификация проводилась для временных рядов с разными мультифрактальными свойствами, полученных генерацией стохастических биномиальных каскадов.

На рис.1 представлены типичные каскады из разных классов: слева каскад с показателем Херста $H = 0.72$, полученный с помощью симметричного бета-распределения с параметром $\alpha = 0.9$, справа каскад, для которого $H = 0.9$ и, соответственно, $\alpha = 3$.

В данном случае каждый класс представляет собой набор сгенерированных временных рядов с одинаковым показателем Херста. В случае, когда весовые коэффициенты являются значениями симметричного бета-распределения, скейлинговая экспонента $\tau(q)$ однозначно определяется значением показателя Херста H, $0.5 < H < 1$. При генерации каскадов значения показателя Херста изменялись в интервале от 0.5 до 1 включая границы с шагом 0.05. Таким образом, обучение моделей производилось на 11 классах.



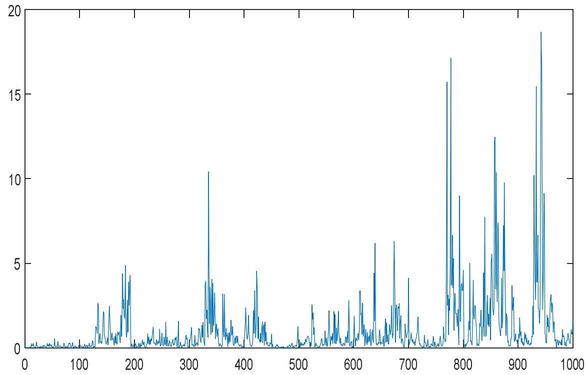


Рисунок 1 - Мультифрактальные каскады:
слева $H = 0.72$, справа $H = 0.9$

В работе для определения принадлежности временного ряда к одному из 11 классов были использованы методы Бэггинг и Случайный лес. В каждом из методов были задействованы ансамбли деревьев решений как классификации, так и регрессии. При использовании регрессионных деревьев решений результатом работы модели является вероятность соответствия мультифрактального каскада заданному классу. Обучение моделей для каждого класса производилось на 5000 примерах временных рядов обучения и проверялось на 500 тестовых.

В таблице представлены средние вероятности определения класса в зависимости от длины мультифрактальных рядов и метода машинного обучения. Результаты показывают, что использование регрессионных деревьев дает существенно большую точность по сравнению с деревьями классификации. Метод Случайный лес показал лучшие результаты, чем метод Бэггинг. Стоит отметить, что Случайный лес проявляет большую стабильность к длине каскадов, что позволяет его применять для коротких временных рядов.

Таблица 1

Средняя вероятность предсказания класса

Длина каскада	Бэггинг		Случайный лес	
	Деревья классификац.	Деревья регрессии	Деревья классификац.	Деревья регрессии
512	0.65	0.79	0.81	0.85
1024	0.66	0.83	0.83	0.88
2048	0.68	0.83	0.84	0.91
4096	0.71	0.89	0.85	0.92
8192	0.75	0.9	0.86	0.93

 ISSN 1562-9945

Выводы

В работе был проведен сравнительный анализ классификации мультифрактальных стохастических временных рядов с использованием мета-алгоритмов на основе деревьев решений. В качестве входных данных использовались биномиальные мультипликативные стохастические каскады. Предложен алгоритм генерации каскадов с заданными мультифрактальными свойствами на основе бета-распределения.

Для классификации рядов использовались методы Бэгиннг и Случайный лес. В каждом методе были задействованы ансамбли деревьев решений, как классификации, так и регрессии. Наилучшие результаты были получены с использованием деревьев регрессии. Полученные результаты могут быть использованы для практических применений, связанных с кластеризацией или кластеризацией реальных временных рядов с фрактальными свойствами.

ЛИТЕРАТУРА

1. Esling P., Agon C.: Time series data mining, ACM Computing Surveys, vol.46, no.1, (2012).

2. Ben D: Feature-based time-series analysis, https://arxiv.org/abs/1709.08055, last accessed 2018/03/03.

3. Bagnall A., Bostrom A., Large J., Lines J.: Simulated Data Experiments for Time Series Classification Part 1: Accuracy Comparison with Default Settings. https://arxiv.org/abs/1703.09480v1, last accessed 2018/03/03.

4. Korus L., Piyrek M.: Compound method of time series classification. Nonlinear Analysis: Modelling and Control, 20(4), 545–560 (2015).

5. Alghawli A., Kirichenko L.: Multifractal Properties of Bioelectric Signals under Various Physiological States. Information Content & Processing International Journal 2(2), 138-163 (2015).

6. Andrй 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).

7. Symeonidis S.: Sentiment analysis via fractal dimension. Proceedings of the 6th Symposium on Future Directions in Information Access, 48-50 (2015).

8. S.P. Arjunan, D.K. Kumar, G.R. Naik A machine learning based method for classification of fractal features of forearm sEMG using Twin Support Vector Machines. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, 4821-4 (2010).

9. R.H.Riedi. Multifractal processes, in Doukhan P., Oppenheim G., Taqqu M.S. (Eds.), Long Range Dependence: Theory and Applications: Birkhuser. -2002. -P. 625–715.

10. L. Kirichenko, T. Radivilova, E. Kayali: Modeling telecommunications traffic using the stochastic multifractal cascade process. Problems of Computer Intellectualization, 55–63 (2012).

11. Breiman L. Bagging predictors. Machine Learning. (1996),24 (2),P.123–140.

12. Breiman L. Random Forests. Machine Learning. (2001), 45 (1), P.5–32.

13. Davy Cielen, Arno Meysman, Mohamed Ali: Introducing Data Science: Big Data, Machine Learning, and more, using Python tools, Manning Publications (2016).

Міністерство освіти і науки України

Системні технології

3 (116) 2018

Регіональний міжвузівський збірник наукових праць

Засновано у січні 1997 року.

У випуску:

– МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ СКЛАДНИХ СИСТЕМ

– ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНО-УПРАВЛЯЮЧІ СИСТЕМИ

Системні технології. Регіональний міжвузівський збірник наукових праць. – Випуск 3 (116). - Дніпро, 2018. – 186с.
ISSN 1562-9945.

Збірник друкується за рішенням Вченої Ради Національної металургійної академії України від 22.01.2018 р., № 1

Адреса редакції: 49635, Дніпро, пр. Гагаріна, 4
Національна металургійна академія України,
кафедра Інформаційних технологій та систем.
Тел. +38(056)7135256
E-mail: st.nmetau.edu.ua
http://st.nmetau.edu.ua

Системні технології

ЗБІРНИК НАУКОВИХ ПРАЦЬ

Випуск 3 (116)

Головний редактор: д.т.н., проф., О.Г. Величко
Шеф-редактор: д.т.н., проф. О.І. Михальов
Комп'ютерна верстка та коректура: к.т.н., доц. К.Ю. Островська

Здано до набору 14.03.2018. Підписано до друку 16.03.2018.
Формат 60х84 1/16. Друк - різограф. Папір типограф.
Умов. друк арк. – 13,23. Обл.–видавн. арк. – 11,625.
Тираж 300 прим. Замовл. – 03/18

Національна металургійна академія України,
кафедра Інформаційних технологій та систем: ІВК «Системні технології»
49005, Дніпро, а/с 493
http://st.nmetau.edu.ua
Свідоцтво про державну реєстрацію друкованого засобу масової інформації:
Серія КВ № 8684 від 23 квітня 2004рік