Local approximation models for human physical activity classification

D. A. Anikeyev[1], G. O. Penkin[2] and V. V. Strijov[3]

[1]Moscow Institute of Physics and Technology, dmitriy.anikeyev@phystech.edu
[2]Lomonosov Moscow State University, penkin.gr@gmail.com
[3]Moscow Institute of Physics and Technology, Dorodnicyn Computing Centre of the Russian Academy of Sciences, strijov@phystech.edu

Abstract: The research is devoted to the time series classification. The time series is measured by an accelerometer of a wearable device. A class of physical activity is defined by its feature description of a time segment. To construct this description the authors propose to use parameters of various approximation splines (algebraic, smoothing, adaptive regression, or spline with dynamic nodes). The logistic regression is used as a classifier. It delivers desired quality of the activity recognition. The authors analyse the space of the local approximation parameters. Classification accuracy depends on the method of this space construction. The computational experiment finds the optimal approximation parameters and parameters of the classifier.

Keywords: local approximation model; time series; classification; splines; feature space

Acknowlegments

The work was partially supported by the Russian Foundation for Basic Research (project 16-07-01158).

References

[1] Karasikov, M. E., and V. V. Strijov. 2008. Feature-Based Time-Series Classification [Klassifikatsiya vremennykh ryadov v prostranstve parametrov porozhdayushchikh modeley]. *Informatika i ee primeneniya* [*Informatics and applications*] 10(4):128–138.

[2] Farmer, J. D., and J. J. Sidorowich. 1987. Predicting chaotic time series. *Physical Review Letters* 59:945–848.

[3] Esling, P., and C. Agon. 2012. Time series data mining. *ACM Computing Survey* 45(1):1–34.

[4] Fu, C. 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence* 24:164–181.

[5] Myasnikov, L. L., and E.N. Myasnikova. 1970. Avtomaticheskoe raspoznavanie zvukovykh obrazov [Automatic speech recognition]. Moscow. *Energiya* [*Energy*]. 183 p.

[6] Tsay, R. S. 2010. Analysis of financial time series. New Jersey. *Wiley.* 712 p.

[7] Coull. B. A., and J. Staundenmayer. 2004. Self modeling regression for multivariate curve data. *Statistica Sinica* 14:695-711.

[8] Fatma, Y. 2011. *CMARS: A new contribution to nonparametric regression with multivariative adaptive regression splines supported by continuous optimisation* Lambert. 212 p.

[9] Istomin, I. A., and O. L. Kotlyarov. 2005. K probleme obrabotki vremennykh ryadov: rasshirenie vozmozhnostey metoda lokal'noy approksimatsii posredstvom singulyarnogo spektral'nogo analiza [The problem of processing time series: Extending possibilities of the local approximation method using singular spectrum analysis]. *Teoreticheskaya i matematicheskaya fizika* [*Theoretical and Mathematical Physics*] 142(1):148–160.

[10] Tselykh, V. R. 2012. Mnogomernye adaptivnye regressionnye splayny [Multivariate adaptive regression splines]. *Mashinnoe obuchenie i analiz dannykh* [*Machine learning and data analysis*] 1(3):272–278.

[11] Dette, H., V. B. Melas, and A. Pepelyshev. 2011. Optimal design for smoothing splines *Annals of the Institute of Statistical Mathematics* 63(5):981–1003.

[12] Gholizadel, S., and E. Salajegheh. 2009. Optimal design of structures subjected to time history loading by swarm intelligence and an advanced metamodels. *Computer methods in applied mechanics and engineering* 198:2936–2949.

[13] Kwapisz, J. R., G. M. Weiss, and S. A. Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12(2):72-82.

[14] Kuznetsov, M. P., and N. P. Ivkin. 2015. Algoritm klassifikatsii vremennykh ryadov akselerometra po kombinirovannomu priznakovomu opisaniyu [Time series classification algorithm using combined feature description]. *Mashinnoe obuchenie i analiz dannykh* [*Machine learning and data analysis*] 1(11):1471–1483.

[15] Vasko, K., and H. Toivonen. 2002. Estimating the number of segments in time series data using permutation tests. *International Conference on Data mining* 466-473.

Alignment of Ordered Set Cartesian Product*

A. V. Goncharov[1], V. V. Strijov[2]

Abstract: The work is devoted to the study of metric methods for analyzing objects with complex structure. It proposes to generalize the dynamic time warping method of two time series for the case of objects defined on two or more time axes. Such objects are matrices in the discrete representation. The DTW method of time series is generalized as a method of matrices dynamic alignment. Paper proposes a distance function resistant to monotonic nonlinear deformations of the Cartesian product of two time scales. The alignment path between objects is defined. An object is called a matrix in which the rows and columns correspond to the axes of time. The properties of the proposed distance function are investigated. To illustrate the method, the problems of metric classification of objects are solved on model data and data from the MNIST dataset.

Key words: distance function; dynamic alignment; distance between matrices; nonlinear time warping; space-time series

BIBLIOGRAPHY

[1] *Hill, N.J., T.N. Lal, M. Schroder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Scholkopf, A. Kubler and N. Birbaumer.* 2006. Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 14 (2): 183–186.

[2] *Sakoe, H. and S. Chiba.* 1971. A dynamic programming approach to continuous speech recognition. Proceedings of the Seventh International Congress on Acoustics. 3: 65–69.

*This work was supported by the RFBR (projects 19-07-1155, 17-07-1574).

[1]Moscow Institute of Physics and Technology, alex.goncharov@phystech.edu

[2]Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control"of the Russian Academy of Sciences, strijov@ccas.ru

Классификация физической активности человека с помощью локальных аппроксимирующих моделей*

Д. А. Аникеев[1], Г. О. Пенкин[2], В. В. Стрижов[3]

Аннотация: Исследуется проблема классификации временных рядов акселерометра мобильного телефона. Классу физической активности соответствует сегмент временного ряда. Сегменту сопоставляется его признаковое описание. Оно порождается аппроксимирующим сплайном. Элементами вектора признаков являются коэффициенты при базисных функциях сплайнов. Вычислительный эксперимент находит оптимальные параметры аппроксимации и параметры модели классификации согласно максимуму правдоподобия логистической модели классификации.

Ключевые слова: временные ряды; классификация; сплайн; локальная аппроксимация; признаковое пространство

1 Введение

Цель работы заключается в решении задачи классификации временных рядов сложной структуры [1, 2], для которых признаковое описание не задано явно. Такие временные ряды встречаются в задачах классификации звуковых сигналов [3], данных с акселерометра [4, 5], построения прогностических финансовых моделей [6]. В [7, 8] дается обзор методов анализа временных рядов за последнее десятилетие. Задача построения признакового пространства требует выбора адекватной гипотезы порождения временного ряда. Многие из исследуемых временных рядов описываются моделью авторегрессии [9, 10] или моделью анализа сингулярного спектра.

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 16-07-01158) и правительства Российской Федерации (соглашение № 05.Y09.21.0018).

[1]Московский физико-технический институт, dmitriy.anikeyev@phystech.edu

[2]Московский государственный университет им. М. В. Ломоносова, penkin.gr@gmail.com

[3]Московский физико-технический институт, Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, strijov@ccas.ru

В данной работе изучается классификация сегментов временных рядов по их локальному описанию. Под локальным описанием будем понимать параметры сплайна, аппроксимирующего сегмент ряда [11–13].

Предполагается, что класс физической активности описывается уникальной аппроксимирующей кривой, моделью, которая задана суперпозицией параметрических функций. В [1, 5] показано, что использование параметров аппроксимирующих моделей в качестве признакового описания сегмента дает ожидаемое качество классификации физической активности.

При этом остается открытой проблема использования значения параметров локальных аппроксимирующих моделей. Точность классификации зависит от этих параметров. Требуется найти оптимальные параметры локальных моделей. Построение моделей, использующих в качестве признакового описания параметры локальных моделей, описано в [14].

В данной работе для решения задачи классификации временных рядов поставлена задача поиска оптимальной локальной аппроксимирующей модели. Сегмент временного ряда аппроксимируется сплайном (алгебраическим, сглаживающим, адаптивным регрессионным или сплайном с динамическими узлами). При помощи логистической регрессии строится отображение из вектора признаков, состоящего из коэффициентов сплайна, в метку класса.

В качестве прикладной задачи рассматривается классификация физической активности человека по данным с акселерометра. В вычислительном эксперименте сравнивается точность классификации в пространствах признаков, построенных различными локальными моделями, найдены оптимальные параметры таких моделей.

2 Постановка задачи классификации сегментов временного ряда

Поставим задачу построения признакового пространства, описывающего сегменты временных рядов с целью их классификации. Задачу сформулируем в терминах построения суперпозиции функций. Задан временной ряд, полученный в результате измерений акселерометра. Для построения выборки $\mathfrak{D}$ ряд разбивается на сегменты. Сегмент временного ряда — конечная последовательность значений $\mathbf{x}(t) = [x_1, \ldots, x_p]^\mathsf{T}$ в моменты времени $\mathbf{t} = [t_1, \ldots, t_p]^\mathsf{T}$. Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, где $y \in \mathbb{Y}$ — метка класса из конечного множества. Требуется найти оптимальный классификатор $f(\mathbf{x}_i)$ из условия

$$\hat{f} = \arg\min_f \mathrm{CV}\big(f, \mathfrak{D}\big), \text{ где } \mathrm{CV}(f, \mathfrak{D}) = \frac{1}{R}\sum_{r=1}^{R}[f(\mathbf{x}_i) \neq y_i] \text{ при } \mathbf{x}_i \in \mathfrak{D} \setminus \mathfrak{D}_r.$$

Выборка $\mathfrak{D}$ случайно разбивается R раз на контрольную $\mathfrak{D}_r$ и тестовую $\mathfrak{D} \setminus \mathfrak{D}_r$. Функцией ошибки выступает число неверно классифицированных объектов $\mathbf{x}_i$ на тестовой

выборке. Предлагается представить f в виде суперпозиции функций $f = g\big(a(t)\big)$. Параметризованное отображение

$$a : (\mathbf{t}, \mathbf{b}) \mapsto \mathbf{x}$$

отображает сегмент временного ряда $\mathbf{x} \in \mathbf{X}$ в пространство параметров аппроксимирующего сплайна, а

$$g : (\mathbf{b}, \mathbf{w}) \mapsto y$$

является логистической регрессией и отображает признаковое описание в метку класса. Вектор $\mathbf{b}$ назовем вектором параметров аппроксимирующей модели, а $\mathbf{w}$ — вектором параметров классификатора g. Оптимальные параметры $(\mathbf{b}, \mathbf{w})$ находятся минимизацией ошибок:

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \ \|a(\mathbf{t}, \mathbf{b}) - \mathbf{x}\|_2^2, \tag{1}$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathrm{CV}\big(f(a, \hat{\mathbf{b}}, \mathbf{w}), \mathfrak{D}\big),$$

где a — сплайн, аппроксимирующий сегмент временного ряда, а $\|\cdot\|_2^2$ — квадратичная ошибка аппроксимации.

3 Выбор локальных аппроксимирующих моделей

Рассмотрим различные методы построения признакового пространства. Зафиксируем сегмент $\mathbf{x} = [x_1, .., x_p]^\mathsf{T}$ временного ряда, порожденный измерениями, получаемыми с носимого устройства в моменты времени $t_1, \ldots, t_p$.

3.1 Алгебраический сплайн

Разобьем сегмент $[t_1, t_p]$ на K равных отрезков. Зададим $\{h_1(t), ..., h_N(t)\}$ — базисные функции сплайна. Гладким сплайном называется функция $a(t)$. Она непрерывна и имеет непрерывные производные вплоть до порядка, называемого гладкостью сплайна, причем на каждом из K отрезков

$$a_k(t) = \sum_{n=1}^{N} \theta_{k,n} \cdot h_n \left(t - \frac{t_k - t_{k-1}}{2} \right), \ k \in \{1, \ldots, K\}, \tag{2}$$

где коэффициенты при базисных функциях $\theta_{k,n} \in \mathbb{R}$.

Оптимальный с точки зрения аппроксимации сплайн ищется из условия (1). В разновидностях сплайнов, аппроксимирующих функцию, накладываются дополнительные условия на функцию. Матрица $\mathbf{\Theta}_{K \times N}$ векторизуется, образуя элемент признакового пространства. Для алгебраического сплайна порядка N базисными функциями являются мономы $h_n = t^n, \ n \in \{1, \ldots, N\}$.

В качестве параметров модели выступают

$$\mathbf{b} = [\mathbf{\Theta}_{K \times N}, N, K]. \tag{3}$$

По каждому сегменту строится $K \times (N+1)$ параметров. В данной работе рассматриваются только квадратичный и кубический сплайны.

3.2 Сглаживающий сплайн

Сглаживающий сплайн минимизирует на отрезке $[t_{\min}, t_{\max}]$ одновременно квадратичную невязку и производную $a(t)$ порядка m. В данной работе используется вторая производная и модель вида (2). Таким образом, минимизируется линейная комбинация

$$(1-\lambda)\sum_{k=1}^{K}\big(x(t_k)-a(t_k)\big)^2+\lambda\int\limits_{t_{\min}}^{t_{\max}}\big(D^2f(t)\big)^2dt\ \to\ \min_{\theta},$$

где λ — параметр гладкости, N — число базисных функций, K — число узлов. В качестве базисных функций выступают мономы не выше заданной степени. Параметрами аппроксимирующей модели являются

$$\mathbf{b}=[\mathbf{\Theta}_{K\times N},\lambda,K,N]. \tag{4}$$

3.3 Адаптивные регрессионные сплайны

В одномерном случае адаптивные регрессионные сплайны (MARS) выражаются через кусочно-линейные базисные функции

$$(\tau-t)_+=\begin{cases}\tau-t, & \text{если } \tau>t,\\ 0 & \text{иначе;}\end{cases}\qquad (t-\tau)_+=\begin{cases}t-\tau, & \text{если } \tau<t,\\ 0 & \text{иначе.}\end{cases}$$

Используемые для аппроксимации базисные функции имеют вид

$$h_j(t)=\prod_{k=1}^{K_j}\left((-1)^{s(j)}(t-\tau_{k_j})\right)_+,$$

где K_j — общее число усеченных линейных функций в m-й базисной функции. На каждом из M шагов алгоритма в множество базисных функций (изначально содержащем одну) добавляется функция вида

$$h_m(t)=\hat{C}h_j(t)(\tau_k-t)_+ + \hat{C}'h_j(t)(t-\tau_k)_+,$$

такая что ошибка (1) минимальна. Таким образом, модель имеет вид

$$a(t)=c_0+\sum_{j=1}^{M}c_jh_j(t)+\varepsilon, \tag{5}$$

где ε — невязка.

В результате добавления модель приобретает избыточное число базисных функций. Необходимо удалить из множества базисных функций те, удаление которых вносит наименьший вклад в увеличение ошибки (1). Параметрами данной модели выступают

$$\mathbf{b} = [\mathbf{H}, M]. \tag{6}$$

Оптимальное число базовых функций

$$\hat{M} = \underset{M\in\{1,\dots,M_{\max}\}}{\arg\min} \ \mathrm{GCV}(M)$$

определяется с помощью критерия обобщенного скользящего контроля

$$\mathrm{GCV}(M) = \frac{1}{R}\sum_{i=1}^{R}\big(x(\tau_i) - a_M(\tau_i)\big)^2 / \big(1 - C(M)/R\big)^2,$$

где R — размер выборки, $C(M)$ — оценка штрафов модели,
$C(M) = \mathrm{trace}\big(\mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}\big) + 1$, в которой $\mathbf{H}$ — матрица с элементами $h_j(\tau_i)$.

3.4 Сплайны с динамическими узлами

Пусть временной сегмент $[t_1, t_p]$ разбит на несколько отрезков $[t_1, t_{n_1}]$, $[t_{n_1}, t_{n_2}], \dots,$ $[t_{n_{K-1}}, t_p]$. Рассмотрим аппроксимацию сегмента временного ряда на каждом отрезке $[t_{n_{k-1}}, t_{n_k}]$ с помощью полинома степени N. Тогда $K \times (N+1)$ коэффициентов матрицы $\mathbf{\Theta}_{K\times N}$ образуют элемент признакового пространства. Предлагается определять концы отрезков последовательно, используя критерий Фишера. Метод заключается в последовательном определении t_{n_k} ($t_{n_0} = t_1$). Положим текущую длину отрезка $l = 1$. Построим полином $y_k^l(t)$ соответствующей степени, минимизирующий RSS на отрезке $[t_{n_k}, t_{n_k+l}]$. Полученные значения $x(t)$ и $y_k^l(t)$ в целых точках на этом отрезке образуют две выборки X и Y. Отношение дисперсий

$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$$

определяет схожесть дисперсий данных выборок. Если дисперсии выборок близки, увеличиваем число l на 1 (кривая хорошо аппроксимирует временной ряд). Формально определяем критерий остановки с помощью p-значения. Если $p(F)$ ниже уровня значимости α — определяем $n_{k+1} = n_k + l$. Продолжим получать последовательность узлов t_{n_k}, пока сегмент не закончится. Число $\alpha \in [0; 0,5]$ будем интерпретировать как параметр метамодели.

Поиск аппроксимирующего сплайна сведем к задаче условной минимизации среднеквадратичного отклонения (1) при соответствующих условиях гладкости функции

$$\frac{\partial^j y_{k+1}(t_{n_k})}{\partial t^j} = \frac{\partial^j y_k(t_{n_k})}{\partial t^j} \quad \text{для всех} \quad k < K, \quad n < N.$$

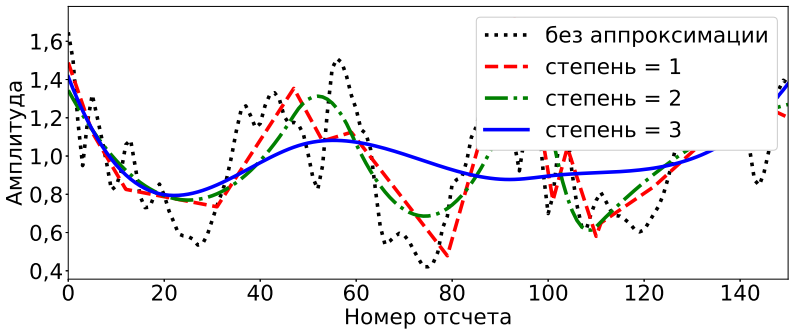


Рис. 1: Аппроксимация сплайном с динамическими узлами

Также рассматривается аппроксимация при условиях равенства только значений сплайна на концах смежных отрезков (кусочно-гладкая). Результаты аппроксимации при параметре $\alpha = 0,07$ и различных степенях полинома показаны на рис. 1.

Критерий Фишера устанавливает разное число узлов на разных сегментах. Способ построения признакового пространства в таком случае заключается в дополнении всех векторов нулевыми компонентами до вектора одинаковой размерности. В качестве альтернативного метода рассмотрим следующую перестановку компонент получившегося вектора.

Назовем $\|\mathbf{y} - \mathbf{g}\|_2^2$ относительным изменением импульса на сегменте, здесь $\mathbf{g} = (0; 0; 9,8)\frac{\text{м}}{\text{с}^2}$ — ускорение свободного падения. Отсортируем отрезки по убыванию относительных изменений импульса, запишем в соответствующем порядке параметры сплайнов и дополним все векторы нулевыми компонентами до векторов одинаковой размерности. Поскольку производится классификация данных с акселерометра, такая сортировка позволяет соотносить между собой отрезки времени с наибольшими изменениями импульса датчика для различных сегментов.

Параметрами модели в этом случае являются

$$\mathbf{b} = [\mathbf{\Theta}_{K\times N}, \alpha, K, N]. \tag{7}$$

3.5 Альтернативные локальные модели

Рассмотрим некоторые альтернативные методы построения признакового пространства.

Дискретное преобразование Фурье. В качестве признакового описания временного ряда берутся коэффициенты прямого преобразования Фурье

$$\theta(\mathbf{x}) = [\theta_1, \ldots, \theta_{2p}], \quad \theta_{2k-1} + i\theta_{2k} = \sum_{j=1}^{p} x^j \exp\left(-\frac{2\pi i}{t} kj\right).$$

Тогда обратное преобразование задает аппроксимацию

$$a(x^k) = \frac{1}{t}\sum_{j=1}^{t} (\theta_{2j-1} + i\theta_{2j}) \exp\left(-\frac{2\pi i}{t} kj\right).$$

Коэффициенты преобразования, образующие признаковое пространство, отыскиваются с помощью линейной регрессии временного ряда на столбцы матрицы Фурье. Применение этого метода для классификации рассматривается в статье [1]

Анализ сингулярного спектра. Элементы признакового пространства — сингулярные числа траекторной матрицы $\mathbf{X}_N$, отвечающие за величины различных частот спектра сегмента

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T},\ \mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}. \tag{8}$$

Тогда собственные числа матрицы $\mathbf{X}^\mathsf{T}\mathbf{X}$ образуют вектор признаков $\mathbf{\Lambda}$. Значение N назначается равным математическому ожиданию длины сегмента по критерию, описанному в разделе сегментации. Результаты классификации в таком пространстве описаны в [5].

4 Сегментация временных рядов

Исходные данные представляют собой размеченные несегментированные отрезки произвольной длины. Для построения выборки необходимо их сегментировать: разбить временной ряд $x_1, \ldots, x_L$ на сегменты вида $[x_1, \ldots, x_p]^\mathsf{T}$ длины p. Рассматривается два варианта разбиения квазипериодических рядов на синфазные сегменты. Алгоритмы выделения периодов рассмотрены в [15, 16].

Выделение главных компонент заключается в разложении временного ряда $\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}} + \boldsymbol{\varepsilon}$, где $\hat{\mathbf{x}}$ — тренд, $\tilde{\mathbf{x}}$ — периодическая часть, а $\boldsymbol{\varepsilon}$ — вектор невязок. Построим траекторную матрицу из элементов временного ряда $x_1, \ldots, x_L$:

$$\mathbf{X} = \begin{bmatrix} x_1 & \cdots & x_p \\ \vdots & \ddots & \vdots \\ x_{L-p+1} & \cdots & x_L \end{bmatrix}.$$

По первым собственным значениям матрицы (8)

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T},\ \mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\},$$

восстановим периодическую зависимость $\tilde{\mathbf{x}} = \mathbf{x}_1 + \cdots + \mathbf{x}_d$, где $\mathbf{x}_i = \sqrt{\lambda_i}\mathbf{v}_i(\mathbf{X}\mathbf{v}_i)^\mathsf{T}$. Искомый период ряда $\hat{p} = \arg\min_p \|\mathbf{x} - \tilde{\mathbf{x}} - \hat{\mathbf{x}}\|_2^2$.

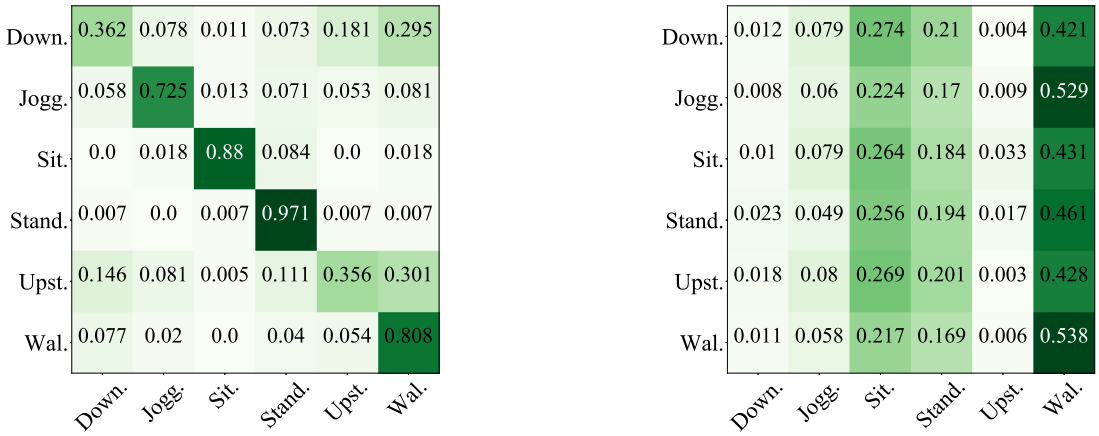
Метод поиска локального максимума временного ряда. Для заданных чисел s и p в каждом отрезке ищется индекс начала сегмента согласно условию $\hat{s} = \arg\max\limits_{j<s} x_j$ и в качестве сегмента берется $[x_{\hat{s}}, \ldots, x_{\hat{s}+p}]^\mathsf{T}$. Для получения синфазных отрезков квазипериодических рядов предлагается использовать параметр s, больший длины периода. Данный метод неустойчив к выбросам, но требует малой вычислительной мощности. Сложность алгоритма линейная по длине исходного сегмента.

5 Вычислительный эксперимент

В качестве вычислительного эксперимента выбрана задача классификации типов физической активности человека по данным с акселерометра.

Данные WISDM. Данные представляют собой размеченные несегментированные трехмерные временные ряды, полученные с датчиков акселерометра. Частота измерений составляла 20 Гц. В выборке представлены классы sitting (225), standing (275), walking (2890), jogging (1631), upstairs (801), downstairs (657). Дробление производилось на временные сегменты длины 50. Для выравнивания сегментов использовался поиск максимального значения за следующие несколько отсчетов, который задает точку начала сегмента.

Сегментация временного ряда проводилась методом поиска максимума с параметрами $m = 20, k = 100$. В качестве классифицирующего алгоритма использована логистическая регрессия из библиотеки sklearn. Результаты классификации показаны на рис. 2.

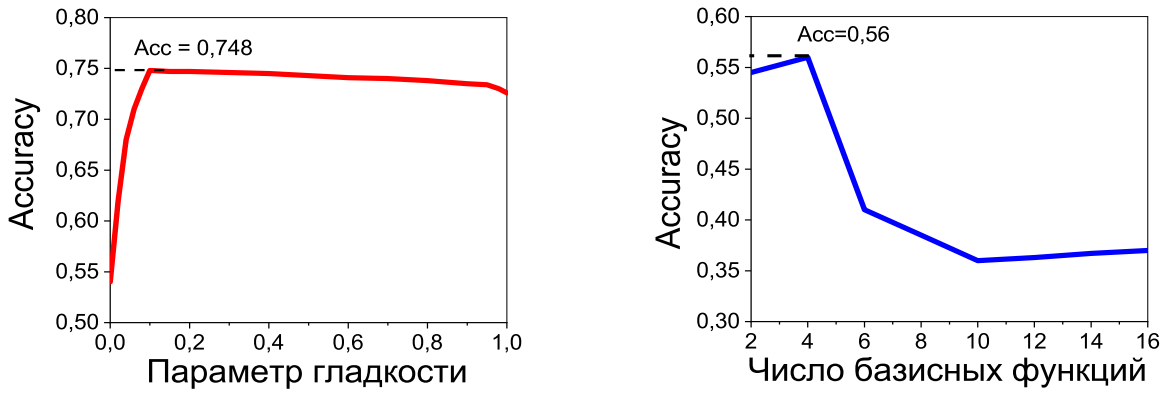


(а) (б)

Рис. 2: Матрица ковариации для кубического (а) и квадратичного (б) сплайна

Точность классификации для квадратичной аппроксимации составляет $0,25$, что ниже результатов остальных исследованных методов аппроксимации. Зависимость точности модели от параметра p для сглаживающего сплайна можно наблюдать на рис. 3а. Как видно из данного графика, максимальная точность достигается при параметре сглаживания $p = 0,25$.

Для построения адаптивного регрессионного сплайна была выбрана библиотека ARESLAB. Для того чтобы размерность пространства параметров была одинакова для всех сегментов, аппроксимация происходила без фазы назад. Параметр такой модели — число базисных функций. График зависимости точности классификации от параметра метамодели показан на рис. 3б. Рассмотрены различные отображения h

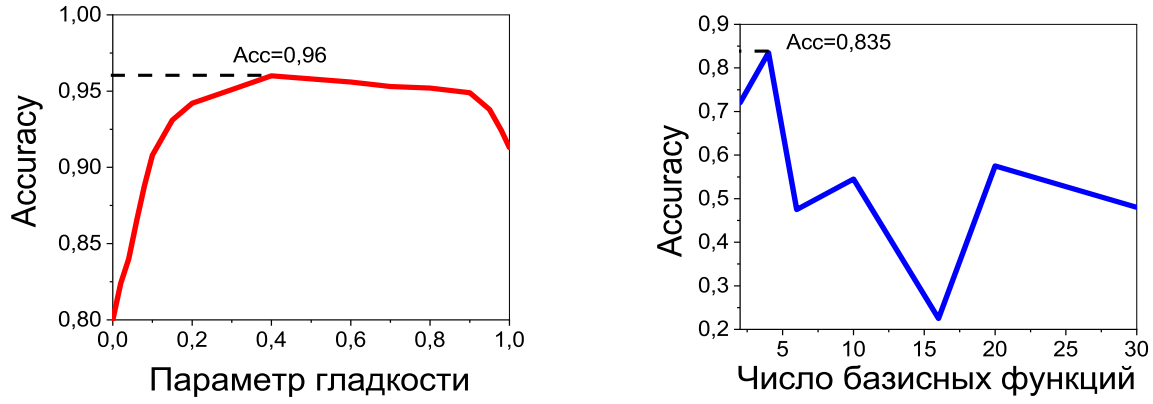


(а) (б)

Рис. 3: Зависимость точности классификации от параметра сглаживания (а) и от числа базисных функций (б)

временных рядов в признаковое пространство. Данные отображения можно сравнить с тождественным, т. е. таким, при котором в качестве матрицы плана выступает матрица значений временного ряда в фиксированные моменты времени. Точность и оптимальные параметры моделей представлены для сравнения в табл. 1.

Данные USC-HAD. Данные представляют собой размеченные временные ряды различной длины с датчиков акселерометра. Даны проекции ускорения на три оси с частотой 100 Гц. В выборке представлены классы walking forward, jumping up, walking left, sitting, walking right, standing, walking upstairs, sleeping, walking downstairs, elevator up, running forward, elevator down.



(а) (б)

Рис. 4: Зависимость точности классификации от параметра сглаживания (а) и от числа базисных функций (б)

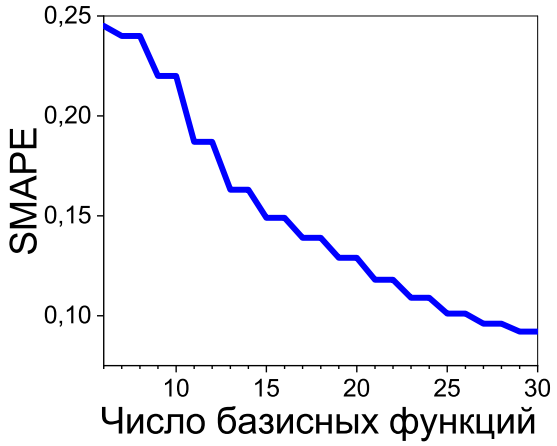


Рис. 5: Зависимость погрешности SMAPE от числа базисных функций

Рассмотрим теперь аппроксимацию гладкими сплайнами с динамическими узлами и кусочно-гладкими, о которых шла речь в разделе 3.4. После аппроксимации ими сегментов имеется возможность менять порядок получаемых признаков. Можно задать его пропорциональным длине кривой между двумя соседними узлами или же пропорционально относительному изменению импульса. Также можно оставить его без изменения.

Все вышеизложенные варианты дают всего шесть различных вариантов (с точностью до степени аппроксимирующего многочлена) задания признакового пространства. На рис. 6 изображено качество классификации в зависимости от значения параметра p.

Таблица 1: Сравнение моделей локальной аппроксимации WISDM и USC-HAD

Аппроксимирующая модель	Оптимальный параметр	Точность
Выборка WISDM		
Quadratic (2)	(3)	0,2540
Cubic (2)	(3)	0,7305
Smoothing spline (2)	$\lambda = 0{,}25$ (4)	0,748
MARS (5)	$M = 4$ (6)	0,56
Выборка USC-HAD		
Quadratic (2)	(3)	0,587
Cubic (2)	(3)	0,926
Smoothing spline (2)	$\lambda = 0{,}4$ (4)	0,960
MARS (5)	$M = 4$ (6)	0,835
Динамические узлы Sort (2)	$K = 2,\ \alpha = 0{,}015$ (7)	0,935
Динамические узлы Unsorted (2)	$K = 2,\ \alpha = 0{,}01$ (7)	0,926

Чтобы результаты классификации для этой выборки можно было сравнить с

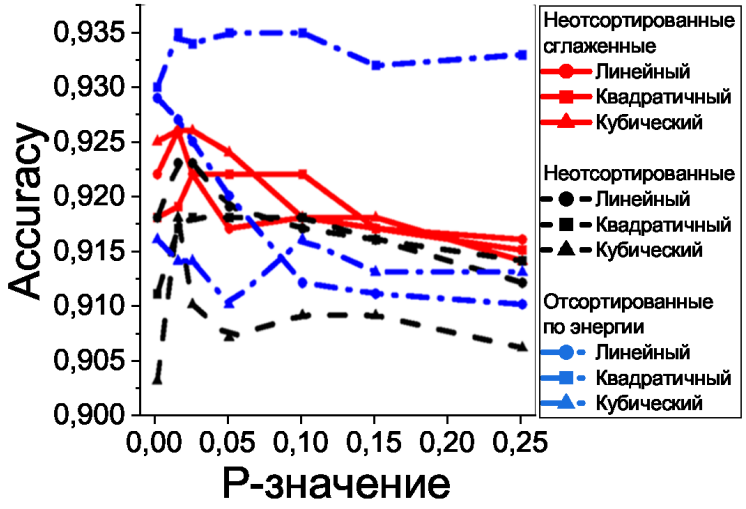


Рис. 6: Точность классификации для сплайнов с динамическими узлами

предыдущей, выделим также шесть классов, по своей природе совпадающих или близких к WISDM. Возьмем классы 1, 6, 7, 8, 9, 10. Будем их для краткости называть walking, running, jumping, sitting, standing, sleeping. Сегментация временного ряда также производилась методом поиска максимума с параметрами $m = 50, k = 200$. Результаты сравнения моделей представлены в табл. 1.

Зависимость погрешности аппроксимации SMAPE от числа базисных функций для регрессионного сплайна показана на рис. 5. На рис. 4б показана зависимость точности классификации от числа базисных функций.

6 Заключение

В работе были описаны различные параметрические модели аппроксимации временных рядов. Проанализирована их эффективность при построении классификатора. Предложен метод классификации временных рядов при помощи построения пространства коэффициентов локальных аппроксимирующих моделей. Сравнивался ряд таких моделей. Квадратичный сплайн оказался непригоден для классификации, остальные отображения показали результат, сравнимый с тождественным отображением. Оптимальные параметры классификатора повышают точность на 3–5%. Предложенный метод позволяет получить классификатор без затрат времени на ручную генерацию признаков.

Список литературы

[1] *Карасиков М. Е., Стрижов В. В.* Классификация временных рядов в пространстве параметров порождающих моделей // Информатика и ее применения, 2008. Т. 10. Вып. 4. С. 121–131.

[2] *Farmer J. D., Sidorowich J. J.* Predicting chaotic time series // Physical Review Letters, 1987. Vol. 59. P. 945–848.

[3] *Мясников Л. Л., Мясникова Е. Н.* Автоматическое распознавание звуковых образов. – Л.: Энергия, 1970. 183 с.

[4] *Kwapisz J. R., Weiss G. M., Moore S. A.* Activity recognition using cell phone accelerometers // ACM SigKDD Explorations Newsletter, 2011. Vol. 12. No. 2. P. 72–82.

[5] *Кузнецов М. П., Ивкин Н. П.* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных, 2015. Т. 1. № 11. С. 1471–1483.

[6] *Tsay R. S.* Analysis of financial time series. – New Jersey: Wiley, 2006. 712 p.

[7] *Esling P., Agon C.* Time series data mining // ACM Computing Survey, 2012. Vol. 45. No. 1. P. 1–34.

[8] *Fu C.* A review on time series data mining // Engineering Applications of Artificial Intelligence, 2011. Vol. 24. P. 164–181.

[9] *Coull B. A., Staundenmayer J.* Self modeling regression for multivariate curve data // Statistica Sinica, 2004. Vol. 14. P. 695–711.

[10] *Yarlikaya F.* CMARS: A new contribution to nonparametric regression with multivariative adaptive regression splines supported by continuous optimisation // Inverse Problems in Science and Engineering, 2012. Vol. 20. No. 3. P. 371–400.

[11] *Истомин И. А., Котляров О. Л.* К проблеме обработки временных рядов: расширение возможностей метода локальной аппроксимации посредством сингулярного спектрального анализа // Теоретическая и математическая физика, 2005. Т. 142. № 1. С. 148–160.

[12] *Целых В. Р.* Многомерные адаптивные регрессионные сплайны // Машинное обучение и анализ данных, 2012. Т. 1. № 3. С. 272–278.

[13] *Dette H., Melas V. B., Pepelyshev A.* Optimal design for smoothing splines // Annals of the Institute of Statistical Mathematics, 2007. Vol. 63. No. 5. P. 981–1003.

[14] *Gholizadel S., Salajegheh E.* Optimal design of structures subjected to time history loading by swarm intelligence and an advanced metamodels // Computer methods in applied mechanics and engineering, 2009. Vol. 198. P. 2936–2949.

[15] *Vasko K., Toivonen H.* Estimating the number of segments in time series data using permutation tests // International Conference on Data Mining. – IEEE, 2006. P. 466–473.

[16] *Motrenko A. P., Strijov V. V.* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2015. Vol. 20. No. 6. P. 1466–1476.