

Порождение признаков с помощью локально-аппроксимирующих моделей*

М. Е. Христолюбов¹

khristolyubov.me@phystech.edu

¹Московский физико-технический институт

В работе рассматривается многоклассовая классификация временных рядов. Классификация производится методом порождения признаков с помощью локально аппроксимирующих моделей. Предполагается, что временные ряды измерений содержат кластеры в пространстве описаний временных рядов. Предлагается разбить временные ряды на сегменты, а в качестве признаков использовать параметры моделей, обученных на этих сегментах. Исследуется оптимальный способ сегментации и информативность порожденных признаков.

Ключевые слова: *временной ряд; многоклассовая классификация; сегментация временных рядов; локально аппроксимирующая модель*

DOI: 00.00000/00000000

1 Введение

В статье изучается задача идентификации движений человека по временным рядам. В дополнении к этому исследуется возможность выделения атрибутивных паттернов, которые могут быть использованы для определения пола или личности субъектов данных в дополнение к их деятельности. Классификация временных рядов находит широкое применение в сфере здравоохранения.

Одним из подходов к классификации временных рядов являются метрические алгоритмы, использующих метрику DTW [4], так как DTW устойчива к локальным смещениям временного ряда. Этот способ сравнения временных рядов в задачах классификации показывает себя лучше, чем использование евклидова расстояния между временными рядами.

Другим подходом является отображение исходного временного ряда в некоторое пространство признаков, так как исходное описание временного ряда (последовательность значений, снятых с определенной частотой) нельзя использовать в качестве признаков, в связи с тем, что отдельные значения несут мало информации о характерном виде временного ряда. Например, в статье [1] временной ряд аппроксимируется моделью, а признаками являются ее параметры. В качестве аппроксимирующей модели берется модель авторегрессии, а так же собственные числа траекторной матрицы, в случае модели сингулярного спектра. В работе [2] проводится разбиение временных рядов на сегменты фиксированной длины, на которых впоследствии обучается локально-аппроксимирующая модель. Для аппроксимации используется линейная модель, модель авторегрессии и коэффициенты преобразования Фурье. Сравняется алгоритм голосования сегментов и классификация в пространстве параметров распределений признаков их сегментов. В [3] предлагается более разумный способ сегментации, а так же применяется аппроксимация сплайнами. Еще более общий подход к способу сегментации посредством нахождения главных компонент

*Работа выполнена при финансовой поддержке РФФИ, проекты № №00-00-00000 и 00-00-00001.

траекторной матрицы, рассмотрен в [6]. В [8] сравниваются между собой перечисленные выше подходы.

В [9] и [10] для классификации используются скрытые марковские модели, но особенность этих работ заключается в классификации действий в каждый момент времени, а не в классификации отдельных временных рядов.

В работе исследуется оптимальный способ сегментации и метод выделения некоторых элементарных движений, по признаковому описанию которых можно будет идентифицировать род деятельности человека, соответствующий временному ряду. Предлагается построить набор локально-аппроксимирующих моделей и выбрать наиболее адекватные. Производится построение метрического пространства описаний элементарных движений, посредством сегментации. Новизна работы заключается в исследовании независимости реализаций временного ряда на различных сегментах. Предположительно, выборка не является полностью независимой, а некоторая зависимость между сегментами характеризует физические параметры испытуемого.

Данные для эксперимента представляют собой измерения акселерометра и гироскопа, встроенных в мобильное устройство iPhone 6s, хранящегося в переднем кармане брюк участника. Временные ряды содержат значения ускорения человека и углы ориентации телефона для каждой из 3 осей — всего 6 временных рядов. Метками классов служат: подъем по лестнице вверх, спуск по лестнице вниз, ходьба, бег трусцой, сидение, лежание. Данные собраны с 24 участников, для которых известны рост, вес, возраст и пол. Данные собирались в условиях проведения эксперимента: участникам выдавали телефон и просили выполнять одно из 6 действий.

2 Постановка задачи

Пусть $x \in \mathcal{X}$ — объекты сложной структуры, представленные временными рядами. Рассматривается задача классификации, а именно восстановление зависимости

$$y = f(x),$$

где $y \in Y$ — пространство ответов.

Заданы выборка \mathfrak{D} объектов сложной структуры и ответов

$$\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^m$$

. Задача состоит в нахождении функции f , минимизирующие суммарные потери на выборке \mathfrak{D} , при заданной функции потерь $\mathcal{L} : (X, F, Y) \rightarrow R$, $\mathcal{L}(f(x_i), y_i)$, характеризующая ошибку классификации функции $f \in F$ на элементе x_i .

Пусть

$$S : X \rightarrow \mathcal{S}, S(x) = \{s_j(x)\}_{j=1}^{N(x)}$$

— процедура сегментации, где $s_j(x)$ — сегменты, возможно, различной длины, и $s_1(x) + \dots + s_{N(x)}(x) = x$, где $+$ означает конкатенацию.

Пусть

$$h : \mathcal{S} \rightarrow W = R^n, h(S(x)) = w(x)$$

— процедура построения признакового описания по набору сегментов. Тогда W — пространство признаков, в котором производится классификация временных рядов.

Пусть b — алгоритм многоклассовой классификации:

$$b : W \rightarrow Y$$

Тогда f ищется в множестве F композиций вида

$$f = b \circ h \circ S$$

Функционалом качества является

$$Q(f, \mathfrak{D}) = \frac{1}{|\mathfrak{D}|} \sum_{(x,y) \in \mathfrak{D}} \mathcal{L}(f(x), y)$$

Для каждой пары (h, S) можно найти оптимальное значение вектора $\hat{\mu}$ параметров классификатора $b(w(x), \mu)$, минимизирующего функционал качества:

$$\hat{\mu} = \arg \min_{\mu} Q(b \circ h \circ S, \mathfrak{D})$$

57 Оптимальный метод обучения, задающийся алгоритмом сегментации S и способом за-
58 дания пространства признаков h , определяется по скользящему контролю

$$f_{h,S}^* = \arg \min_{h,S} \widehat{CV}(f_{h,S}, \mathfrak{D}),$$

59 где $\widehat{CV}(f, \mathfrak{D})$ — внешний контроль качества метода обучения f , $\mathfrak{D} = \mathfrak{L} \sqcup \mathfrak{E}$:

$$\widehat{CV}(\mu, \mathfrak{D}) = \frac{1}{r} \sum_{k=1}^r Q(f^*(\mathfrak{L}), \mathfrak{E})$$

В качестве функционала качества используется

$$Q(f, \mathfrak{L}) = \frac{1}{|\mathfrak{L}|} \sum_{(x,y) \in \mathfrak{L}} |\{(x, y) \in \mathfrak{L} | f(x) = y\}|$$

, а в оценке точности классификации объектов класса $c \in Y$ используется модифицированный функционал качества

$$Q_c(f, \mathfrak{L}) = \frac{1}{|\mathfrak{L}|} \sum_{(x,y) \in \mathfrak{L}} \frac{|\{(x, y) \in \mathfrak{L} | f(x) = y = c\}|}{|\{(x, y) \in \mathfrak{L} | y = c\}|}$$

60 2.1 Название параграфа

61 Разделы и параграфы, за исключением списков литературы, нумеруются.

62 3 Заключение

63 Желательно, чтобы этот раздел был, причём он не должен дословно повторять ан-
64 нотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы
65 остались открытыми.

66 Литература

67 *Поступила в редакцию 00.00.0000*

Machine Learning and Data Analysis journal paper template*

F. S. Author¹, F. S. Co-Author², and F. S. Name^{1,2}

khristolyubov.me@phystech.edu

¹Organization, address; ²Organization, address

This is the template of the paper submitted to the journal “Machine Learning and Data Analysis”. The title should be concise and informative. Titles are often used in information-retrieval systems. Avoid abbreviations and formulae where possible. A concise and factual abstract is required. **Background:** One paragraph about the problem, existent approaches and its limitations. **Methods:** One paragraph about proposed method and its novelty. **Results:** One paragraph about major properties of the proposed method and experiment results if applicable. **Concluding Remarks:** One paragraph about the place of the proposed method among existent approaches. **Keywords:** *keyword; keyword; more keywords, separated by “;”*

DOI: 00.00000/00000000

References

- [1] N. P. Ivkin, M. P. Kuznetsov.. 2015. Time series classification algorithm using combined feature description. . *Machine Learning and Data Analysis* (11):1471–1483.
- [2] V. V. Strijov, M. E. Karasikov. 2016. Feature-based time-series classification *Informatics* doi: <http://dx.doi.org/10.3114/S187007708007>.
- [3] D.A. Anikeev, G.O. Penkin, V.V. Strijov. 2018. Local approximation models for human physical activity classification // *Informatics* doi: <http://dx.doi.org/10.14357/19922264190106>.
- [4] V.V. Strijov, R.V. Isachenko.. 2016. Metric learning in multiclass time series classification problem. *Informatics and Applications* (10(2)):48–57.
- [5] V.V. Strijov, Andrew Zadayanchuk, Maria Popova.. 2016. Selection of optimal physical activity classification model using measurements of accelerometer. *Information Technologies* (22(4)):313–318.
- [6] Strijov V.V., Motrenko A.P.. 2016. Extracting fundamental periods to segment human motion time series. *Journal of Biomedical and Health Informatics* 20(6):1466 – 1476.
- [7] Strijov V.V., Ignatov A.. 2015. Human activity recognition using quasiperiodic time series collected from a single triaxial accelerometer. *Multimedia Tools and Applications* pages 1–14.
- [8] Isachenko R.V., Bochkarev .., Zharikov I.N., Strijov V.V.. 2018. Feature Generation for Physical Activity Classification. *Artificial Intelligence and Decision Making* 3 : 20-27.
- [9] Dafne van Kuppevelt, Joe Heywood, Mark Hamer, Séverine Sabia, Emla Fitzsimons, Vincent van Hees. 2019. Segmenting accelerometer data from daily life with unsupervised machine learning. *PLOS ONE* doi: <http://dx.doi.org/10.5255/UKDA-SN-8156-3>.
- [10] Andrea Mannini, Angelo Maria Sabatini. 2010. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers *PubMed* doi: <http://dx.doi.org/10.3390/s100201154>.
- [11] Malekzadeh, Mohammad and Clegg, Richard G. and Cavallaro, Andrea and Haddadi, Hamed. 2019. *Mobile Sensor Data Anonymization* pages 49–58. Proceedings of the International Conference on Internet of Things Design and Implementation doi: <http://dx.doi.org/10.1145/3302505.3310068>.

Received January 00, 0000

*The research was supported by the Russian Foundation for Basic Research (grants 00-00-0000 and 00-00-00001).