Порождение признаков с помощью локально-аппроксимирующих моделей*

М. Е. Христолюбов[1]
khristolyubov.me@phystech.edu
[1]Московский физико-технический институт

В работе рассматривается многоклассовая классификация временных рядов акселерометра. Классификация производится методам порождения признаков с помощью локально аппроксимирующих моделей. Предполагается, что точки временного ряда можно разбить на кластеры, соответствующие разным классам. Предлагается выделить квазипериодические сегменты из временных интервалов точек, принадлежащих одному кластеру. В качестве признаков для классификации использовать параметры моделей, обученных на этих сегментах. Исследуется вопрос информативности порожденных признаков и возможность идентификации по ним владельца прибора или его физических параметров.

Ключевые слова: *временной ряд; многоклассовая классификация; кластеризация; сегментация временного ряда; 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 выделение атрибутивных паттернов, которые могут быть использованы для определения личности или физических параметров субъектов данных в дополнении к их деятельности. Классификация временных рядов находит широкое применение в сфере здравоохранения.

Временные ряды являются объектами сложной структуры, требующие предварительной обработки и представления их в удобном для классификации виде. Необходимо отобразить исходный временной ряд в некоторое пространство признаков. Например, в статье [1] временной ряд аппроксимируется моделью, а признаками являются ее параметры. В качестве аппроксимирующей модели берется модель авторегрегрессии, 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. В [3] предлагается более разумный способ сегментации, а так же применяется аппроксимации сплайнами. Еще более общий подход к способу сегментации посредством нахождения главных компонент траекторной матрицы, рассмотрен в [6]. В [8] сравниваются между собой перечисленные выше подходы.

Однако, вышеперечисленные подходы работают только в случае, когда заранее дан временной ряд, соответствующий одному виду деятельности, что невозможно в реальных условиях. Реальные данные представлены временным рядом, для которого в каждый момент времени нужно определить род деятельности. Метод кластеризации точек, соответствующих участкам разной деятельности, с помощью методы главных компонента (SSA,

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

алгоритм гусеница [12]) рассмотрен в [11]. На участках, содержащих точки одного кластера, уже можно применять описанные выше методы. Другим подходом к классификации точек временного ряда на основе нейросетей рассмотрены в [9] и [13].

В работе исследуется оптимальный способ кластеризации и сегментации, способ выделения некоторых элементарных движений, по признаковому описанию которых можно будет идентифицировать род деятельности человека. Предлагается построить набор локально-аппроксимирующих моделей и выбрать наиболее адекватные. Производится построение метрического пространство описаний элементарных движений. 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 и может быть использована для идентификации.

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 ориентацию 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 имеется исходный временной ряд $d = \{d_i\}_{i=1}^{M} \in \mathbb{R}^M$. Предполагается, что он состоит из последовательности сегментов:

$$d = [s_1, \dots s_N],$$

где $s_i \in \mathcal{S}$, $|S|$ — число различных действий (кластеров). Считается, что периоды $|s|$ сегментов различаются незначительно, причем известен максимальный период $|s| \leqslant T$. Тип активности не меняется часто, то есть можно выделить участки временного ряда, соответствующие одному типу активности.

Требуется решить задачу классификации точек ряда:

$$R : \mathcal{I} \to Y,$$

где $\mathcal{I} = \{1, \dots M\}$ — индексы точек ряда, а Y — метки классов.

Предварительно требуется решить задачу кластеризации, то есть нахождения отображения:

$$a : \mathcal{I} \to Z,$$

где Z — множество меток кластеров.

Сопоставление меток кластеров и меток классов проведем посредством классификации временных интервалов $x = \{d_t, \dots d_{t+T}\}$, в которых последовательный набор точек исходного ряда принадлежит одному кластеру: $\forall i \in [t, t+t] : a(d_i) = z_x$. Процедуру выделения из исходного ряда интервалов x посредством кластеризации обозначим $g(d) = \{x_1, \dots x_P\}$

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. Тогда исходная задача классификации представляет собой $R = f \circ g$, где f применяется ко всем интервалам исходного ряда $x \subset d$.

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 функция потерь $\mathscr{L} : (\mathcal{X}, F, Y) \to R$, $\mathscr{L}(f(x_i), y_i)$, характеризующая ошибку классификации функции $f \in F$ на элементе x_i.

Пусть

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

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

Пусть

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

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

Пусть b — алгоритм многоклассовой классификации:

$$b : W \to Y$$

Тогда f ищется в множестве F композиций вида

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

Функционалом качества является

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

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

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

Оптимальный метод обучения, задающийся алгоритмом сегментации S и способом задания пространства признаков h, определяется по скользящему контролю

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

где $\widehat{CV}(f, \mathfrak{D})$ — внешний контроль качества методы 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

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

2.1 Название параграфа

Разделы и параграфы, за исключением списков литературы, нумеруются.

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

Поступила в редакцию 00.00.0000

Machine Learning and Data Analysis journal paper template*

***F. S. Author*[1], *F. S. Co-Author*[2], *and F. S. Name*[1,2]**

khristolyubov.me@phystech.edu

[1]Organization, address; [2]Organization, 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.

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

[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] Grabovoy A.V., Strijov V.V. 2020. Quasiperiodic time series clustering for human activity recognition *Lobachevskii Journal of Mathematics*

[12] D.L. Danilov and A.A. Zhiglovsky. 1997. *Main components of time series: method "Gesenitsa"(St. Petersburg)*

[13] Y.G. Cinar and H. Mirisaee. 2018. Period-aware content attention RNNs for time series forecasting with missing values "*Neurocomputing* 312, 177–186

[14] 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