Порождение признаков с помощью локально-аппроксимирующих моделей.*

Садиев А. А.[1]***, Фатхуллин И. Ф.***[1]***, Мотренко А. П.***[1]***, Стрижов В. В.***[1]
sadiev.aa@phystech.edu, ilyas.fn979@gmail.com
[1]Московский физико-технический институт (МФТИ)

Рассматриваются методы классификации физической активности человека по измерениям акселерометра. Статья посвящена исследованию проблемы порождения признаков с использованием локально-аппроксимирующих моделей. В работе строится набор локально-аппроксимирующих моделей и проверяется корректность гипотезы о простоте выборки для порожденных признаков. Решается задача выбора оптимального способа порождения признаков временного ряда. В контексте данной работе предполагается метод построения метрического пространства описаний элементарных движений.

Ключевые слова: *временной ряд, многоклассовая классификация, локально-аппроксимирующая модель, метрическое пространство.*

1 Введение

Работа посвящена поиску оптимальных признаков для задачи классификации видов физической активности человека. Исследование проводится с целью автоматизации порождения признаков слабоструктурированных данных, таких как временные ряды. Оптимальный выбор признаков должен удовлетворять выборкам временных рядов с различными частотами. Также предлагаемый в данной работе метод должен обеспечивать минимальное расхождение в точности задачи классификации с различными множествами ответов.

Задача оптимального порождения признаков решается различными способами [1–6]. В работе [2] выделяются фундаментальные периоды временных рядов, в [5, 6] внимание уделено сегментации временного ряда различными способами. Также стоит отметить использование сплайнов в порождении признаков временнго ряда [4], в статье [1] предложен новый метод с использованием кубических сплайнов, которые дают гладкую кривую и приемлемое качество аппроксимации. Помимо классических методов применяются нейронные сети, а именно построение нейронной сети оптимальной структуры для решения задачи классификации. В работе [3] используются два алгоритма на нейронных сетях для получения решения задачи классификации.

В данной работе задача решается с помощью построения универсальной стандарта . Он состоит из суперпозиции локально-аппроксимирующих моделей исходной выборки. Предлагаемый метод не дает наилучшую точность среди уже имеющихся спрособов , однако является универсальным для данных с различными параметрами выборок. Однако, был создана библиотека локально-аппроксимирующих моделей, удобно используемая на практике.

Вычислительный эксперимент проводится на данных временных рядов акселерометра WISDM с целью решения задачи классификации.

---

*Работа выполнена при финансовой поддержке РФФИ, проект № 00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт И. О. Консультант: Мотренко А. П.

2 Постановка задачи

Пусть задана выборка $\mathfrak{D} = \mathrm{z}\{(\mathbf{s}_i, y_i)|\ i = 1, ..., m;\ \mathbf{s}_i = [\mathbf{s}_i(1), \ldots, \mathbf{s}_i(T)] \in \mathbf{S} \subset \mathbb{R}^{n\times m}\}$, где $\mathbf{s}_i(t) \in \mathbb{R}^n$, $y_i \in Y$ - пространство ответов, $|Y| = K \in \mathbb{N}$, m - количество элементов в выборке. Поставим задачу многоклассовой классификации временных рядов. Временные ряды являются объектами сложной структуры. Поэтому процесс классификации разбивают на два основных этапа: первый - порождение признакового описания (создание пространства признаков), второй - решение задачи классификации. Формально задача классификации состоит в определении отображения $f : \mathbf{S} \to Y$. Отображение будем искать в виде суперпозиции:

$$f(\mathbf{s}) = g(h(\mathbf{s}), \mathbf{w}) \tag{1}$$

где $h : \mathbf{S} \to \Phi$. $\Phi \subset \mathbb{R}^p$ - пространство признаков, $\mathbf{w}$ - вектор параметров модели.

Чтобы определить качество работы классификатора, задается функция потерь $\mathcal{L}\left(f(\mathbf{s}_i), y_i)\right)$, выражающая величину ошибки классификации отображения f на объекте $\mathbf{s_i}$ данной выборки $\mathfrak{D}$. Таким образом, для решения нашей задачи нужно найти отображение f, минимизирующая суммарную фунцию потерь на выборке $\mathfrak{D}$:

$$\mathbf{y}_{opt} = \arg\min_{\mathbf{y}\in\mathbb{R}^m} \sum_{i=1}^{m} \mathcal{L}(f(\mathbf{s}_i, \mathbf{w})] \tag{2}$$

Функция f, определенная в (1), является суперпозицией отображений $g(., \mathbf{w})$ и $h(s)$. В данной работе исследуются свойства функций вида $h : \mathbf{S} \to \Phi$: она порождает признаковое описание объектов $\mathbf{s}_i$ из данной выборки $\mathfrak{D}$. Есть множество способов определить h, например, с помошью алгоритмов AR, DFT, SSA, SEMOR и т. д. Поэтому будем рассматривать модели $h_j \in \mathcal{H}$, где $j \in \{1, \ldots, r\}$, где r - количество моделей в наборе $\mathcal{H}$. Эти функции создают признаковое описание объекта $\mathbf{s}_i$ (каждая свое), т. е. $h_j(\mathbf{s}_i) = \boldsymbol{\varphi}^{(ij)} = [\varphi_1^{(ij)}, \ldots, \varphi_p^{(ij)}]^T \in \Phi$. Допустим на первом этапе каким-либо образом получено подмножество $\mathcal{P} \subset \mathcal{H}$ алгоритмов из заданного набора. Подмножеству $\mathcal{P}$ соответствует признаковое описание, полученное конкатенацией признаков алгоритмов из $\mathcal{P}$. Тогда на втором этапе имеем классическую задачу многоклассовой классификации:

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{L}[g(\mathcal{P}, \mathbf{w})] \tag{3}$$

В итоге, объединяя два этапа, получаем задачу вида:

$$\mathcal{P}_{opt} = \arg\min_{\mathcal{P}\subset\mathcal{H}} \min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{L}\left[g(\mathcal{P}, \mathbf{w})\right] \tag{4}$$

3 Порождение признаков

Как было описано выше функция $h : \mathbf{S} \to \Phi$, где $\Phi \subset \mathbb{R}^p$ - пространство признаков, порождает ращлчные признаки. Стоит отметить, что мы работаем с сегментом, как с объектом. Приведем какие функции мы использовали:

3.1 Дискретное преобразование Фурье

Берется временной ряд $\mathbf{s}_i \in \mathbf{S}$ и производят сегментацию: получают набор $\{\mathbf{s}_i^k\}_{k=0}^N$, где N - количество полученных сегментов. Далее работаем с одним сегментом как с объектом: применяем к нему дискретное преобразование Фурье

$$f_k = \sum_{n=0}^{N'-1} s_i^k[n] e^{-\frac{2\pi i}{N'}kn}, \qquad (k = 0, \ldots, N'-1) \tag{5}$$

N' - количество элементов в s_i^k. Так как полученные коэффициенты копмлексные, что не понятно как интерпретировать физически, то мы комплексное число представим в полярном виде: компоненты вектора признаков у величится в двое. Таким образом, мы получаем вектор признаков $\mathbf{f} = (f_1 \quad f_2 \dots f_{2N'})$.

3.2 Статистические функции

Берется временной ряд $\mathbf{s}_i \in \mathbf{S}$ и производят сегментацию: получают набор $\{\mathbf{s}_i^k\}_{k=0}^N$, где N - количество полученных сегментов. Далее работаем с одним сегментом как с объектом:

- Среднее значение: $\overline{m}_i^k = 1N' \sum_{n=1}^{N'} \mathbf{s}_i^k[n]$
- Дисперсия: $d_i^k = 1N'^2 \sum_{n=1}^{N'} ((\mathbf{s}_i^k[n])^2 - (\overline{m}_i^k)^2)$
- Абсолютное отклонение $\alpha_i^k = 1N'^2 \sum_{n=1}^{N'} (\mathbf{s}_i^k[n] - \overline{m}_i^k)$

Применяя эти функции к каждому сегменту, мы порождаем признаки.

3.3 Авторегрессия

Берется временной ряд $\mathbf{s}_i \in \mathbf{S}$ и производят сегментацию: получают набор $\{\mathbf{s}_i^k\}_{k=0}^N$, где N - количество полученных сегментов. Далее работаем с одним сегментом как с объектом. Авторегресия учитывает предысторию, что логично использовать. Поэтому это записывается в следущем виде :

$$\mathbf{s}_i^k[t+1] = \sum_{j=1}^{l} w_j s_i^k[t-j+1], \tag{6}$$

где l - количество предыдущих наблюдений ряда (сегмента), $\hat{\mathbf{w}}$ - вектор параметров модели авторегрессии. Формулу (6) перепишем в маричном виде:

$$F^{\alpha\times l} = \begin{pmatrix} s_i^k[t-1] & s_i^k[t-2] & \dots & s_i^k[t-l] \\ s_i^k[t-2] & s_i^k[t-3] & \dots & s_i^k[t-l-1] \\ \dots & \dots & \dots & \dots \\ s_i^k[l] & s_i^k[l-2] & \dots & s_i^k[1] \\ s_i^k[l-1] & s_i^k[l-2] & \dots & s_i^k[0] \end{pmatrix}$$

$$\mathbf{y}^{\alpha\times 1} = \begin{pmatrix} s_i^k[t] \\ s_i^k[t-1] \\ \dots \\ s_i^k[n+1] \\ s_i^k[n] \end{pmatrix}$$

где в роли объектов $\alpha = t - l + 1$ моментов из истории. Тогда чтобы найти вектор параметров данной модели нужно решить следующую задачу минимизации:

$$\hat{\mathbf{w}}_{opt} = \underset{\hat{\mathbf{w}}\in\mathbb{R}^l}{\arg\min} \|F\hat{\mathbf{w}} - y\| \tag{8}$$

Таким образом, полученным вектором параметров мы будем характеризовывать наш сегмент.

4 Базовый вычислительный эксперимент

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

Данные WISD представляют собой трехмерные временные ряды, полученные с датчика акселерометра, причем данные размечены, но не сегментированы. Частота измерений составляет 20 Гц. В данной выборке представлены 6 классов: sitting (225), standing (275), walking (2890), jogging (1631), upstairs (801), downstairs (657).

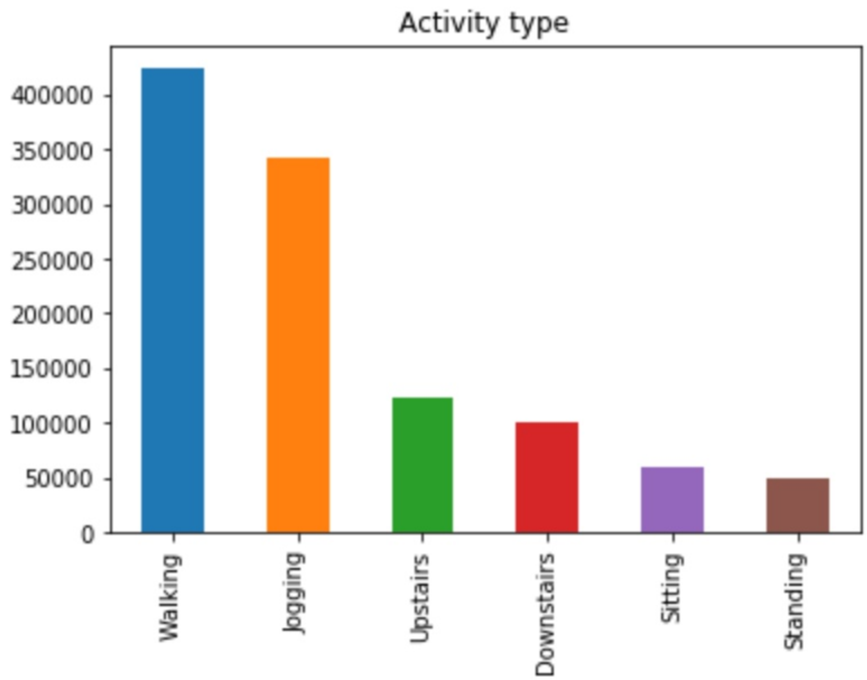

Рис. 1 Количество измерений для каждого класса

Сегментация ряда произдодилось делением на равномерные части (сегменты). Стандартно количество сегментов $n = 200$. В качестве модели классификации рассматривались логистическая регрессия, случайный лес и метод опорных векторов SVM.

Используя приведенные выше алгоритмы порождения признаков, мы смогли создать матрицу объект-признак , и далее использовать ее для классификации данных по 6 классам. Приведем полученный результат:

	all	Jogging	Upstairs	Standing	Walking	Downstairs	Sitting
lr_all_feat_	0.940142	0.990678	0.959557	0.996075	0.966426	0.971332	0.996215
lr_all_feat_sampled_	0.944137	0.993411	0.963412	0.995094	0.969510	0.969931	0.996916
rf_all_feat_	0.961940	0.991028	0.973015	0.997757	0.981566	0.983388	0.997126
rf_all_feat_sampled_	0.968459	0.991799	0.976309	0.998668	0.985912	0.986192	0.998037
svm_all_feat_	0.975608	0.995094	0.988855	0.995023	0.988855	0.988785	0.994603
svm_all_feat_sampled_	0.979673	0.996355	0.992360	0.994673	0.989977	0.991659	0.994323

Рис. 2 Результаты

5 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

[1] Zharikov I. N. Strijov V. V. Isachenko R. V., Bochkarev A. M. Feature generation for physical activity classification. *Artificial Intelligence and Decision Making*, pages 20–27, 2018.

[2] Anastasia Motrenko and Vadim Strijov. Extracting fundamental periods to segment biomedical signals. *IEEE J. Biomedical and Health Informatics*, 20(6):1466–1476, 2016.

[3] В. В. Стрижов А. И. Задаянчук, М. С. Попова. Выбор оптимальной модели классификации физической активности по измерениям акселерометра. *Информационные технологии*, 22(4):313–318, 2016.

[4] Д. А. Аникеев В. В. Стрижов, Г. О. Пенкин. Классификация физической активности человека с помощью локальных аппроксимирующих моделей. *Информ. и её примен.*, 18(1):144–156, 2018.

[5] М. Е. Карасиков В. В. Стрижов. Классификация временных рядов в пространстве параметров порождающих моделей. *Информ. и её примен.*, 10(4):121–131, 2016.

[6] М. П. Кузнецов Н. П. Ивкин. Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию. *Машинное обучение и анализ данных.*, 1(11):1471–1483, 2015.