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
[1]Московский физико-технический институт (МФТИ)

Рассматриваются методы определения вида деятельности человека по измерениям акселерометра. Статья посвящена исследованию проблемы порождения признаков с использованием локально-аппроксимирующий моделей. В работе строится набор локально-аппроксимирующих моделей и проверяется корректность применения гипотезы о простоте выборки для порожденных признаков. Также внимание уделено выбору 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 деятельности человека. Исследование проводится с целью автоматизации порождения признаков слабоструктуированных 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.

Проблема оптимального порождения признаков решается множеством способов: в работе [1] выделяются фундаментальные периоды временных рядов, в [3] [2] внимание уделено сегментации временного ряда различными способами. Также стоит отметить использование сплайнов в порождении признаков временнго ряда [6], в статье [7] 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 [5] используются два алгоритма на нейронных сетях для получения решения задачи классификации.

В данной работе задача решается с помощью построения библиотеки локально-аппроксимирующих моделей исходной выборки. Предлогаемый метод не дает наилучшую точность среди уже имеющихся спрособов , однако является универсальным для данных с различными параметрами выборок.

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

2 Постановка задачи

Пусть задана выборка $\mathfrak{D} = \{(\mathbf{s}_i, y_i)|\ i = 1, ..., m;\ \mathbf{s}_i = [\mathbf{s}_i(1), \ldots, \mathbf{s}_i(T)] \in \mathbf{S} \subset \mathbb{R}^{n\times m}\}$, 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

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

являются объектами сложной структуры. Поэтому процесс классификации разбивают на два основных этапа: первый - порождение признакового описания (создание пространства признаков), второй - сама классификация. Формально задача классификации состоит в определении отображения $f : \mathbf{S} \to Y$. В силу вышесказанного отображение будем искать в виде суперпозиции: $f(\mathbf{s}) = g(h(\mathbf{s}), \mathbf{w})$, 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 $\mathfrak{D}$.

Как было определено выше, функция f является суперпозицией отображений $g(., \mathbf{w})$ и $h(s)$. Рассмотрим подробней функцию $h : \mathbf{S} \to \Phi$: 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:

$$\mathbf{w}_{opt} = \arg\min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{L}[g(\mathcal{P}, \mathbf{w})] \tag{1}$$

В итоге, объединяя два этапа, получаем задачу вида:

$$\mathcal{P}_{opt} = \arg\min_{\mathcal{P} \subset \mathcal{H}} \min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{L}\left[g(\mathcal{P}, \mathbf{w})\right] \tag{2}$$

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

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

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