Локальные методы прогнозирования, поиск метрики*

Максим Христолюбов, В. В. Стрижов
Московский физико-технический институт

В работе решается задача прогнозирования временных рядов. Временной ряд делится на отдельные участки, каждому из которых сопоставляется точка в n-мерном пространстве признаков. Локальная модель рассчитывается в три последовательных этапа. Первый – находит k-ближайших соседей наблюдаемой точки. Второй – строит простую модель, используя только этих k соседей. Третий – используя данную модель, по наблюдаемой точке прогнозирует следующую. Данная работе исследуется оптимальный набор весов во взвешенной метрике для максимизации точности прогнозирования.

Ключевые слова: *временной ряд; классификация; сегментация временного ряда; локально-аппроксимирующая модель, порождение признаков*

1 Введение

В статье изучается задача прогнозирование движений человека по временным рядам акселлерометра. Методы построения прогноза временных рядов делятся на глобальные (использующие всю предысторию ряда) и локальные (используют только её часть).В данной работе рассматривается локальный метод прогнозирования, основанный на алгоритме поиска k ближайших соседей, который был описан в работах Дж. Макнеймса [1] и Ю.И. Журавлева [2]. Новизна работы в том, чтобы искать метрику не в исходном пространстве временных рядов, а отобразить каждый участок временного ряда в пространство признаков.

Временные ряды являются объектами сложной структуры, требующие предварительной обработки и представления их в удобном для сравнения виде. Метод кластеризации точек, соответствующих участкам разной деятельности, с помощью методы главных компонент (SSA, алгоритм гусеница [**?**]) рассмотрен в [5]. Предлагается использовать отображение локального участка ряда в пространство главных компонент для нахождения k ближайших соседей.

2 Постановка задачи

Задан исходный временной ряд $\mathbf{d} = \{d_i\}_{i=1}^M \in \mathbb{R}^M$.

Задача прогнозирования временного ряда состоит в том, чтобы по известному отрезку временного ряда

$$(d_1 \dots d_n)$$

предсказать следующие l его значений:

$$(d_{n+1} \dots d_{n+l})$$

Решается задача построения локального метода прогнозирования временных рядов, основанного на алгоритме "ближайших соседей". В базовом варианте для оценки степени близости объектов предлагается использовать евклидову метрику в исходном пространстве, а после сравнить качество прогноза при модификации алгоритма. В базовом варианте алгоритм "ближайших соседей" состоит из следующих этапов:

*

1. Найти в предыстории среди всех векторов размерности l, составленных из отрезков временного ряда ($\mathbf{f_i} = f_i, \dots, f_{i+l-1}$), k векторов, наиболее похожих на вектор $(f_{nl+1}, f_{nl+2}, \dots, f_n)$.

2. Пусть $(f_{i_1 l+1}, \dots, f_{i_1}), \dots, (f_{i_k l+1}, \dots, f_{i_k})$) — k ближайших соседей для предыстории $(f_{n-l+1}, \dots, f_n)$. Прогноз $(\hat{f}_{n+1}, \dots, \hat{f}_{n+t})$ вычисляется как взвешенное среднее арифметическое этих k векторов.

В модифицированном алгоритме, в отличиие от базового, перед нахождением наиболее похожих, каждый участок временного ряда длинны l отображается в пространство главных компонент в соответствии с алгоритмом гусеница:

0. Поставим в соответствие временному сегменту $f_i \ \dots f_{i+l-1}$ его траекторную матрицу $\mathbf{X}$. Ее сингулярное разложение

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{H}\mathbf{V}^\mathsf{T}, \quad \mathbf{H} = \mathrm{diag}(h_1, \dots h_l).$$

$h_1 \dots h_l$ — собственные числа матрицы $\mathbf{X}^\mathsf{T}\mathbf{X}$. Первые n собственных чисел берется в качестве нового признакового описания $\mathbf{h} = [h_1, \dots h_n]$. Тогда $\rho(\mathbf{f_i}, \mathbf{f_j}) = \rho(\mathbf{h_i}, \mathbf{h_j}) = \sqrt{(\mathbf{h}_i - \mathbf{h}_j)^T (\mathbf{h}_i - \mathbf{h}_j)}$

Для оценки качества алгоритма используется функционал ошибки Symmetric Mean Absolute Percent:

$$SMAPE(\mathbf{f}, \hat{\mathbf{f}}) = \frac{1}{l} \sum_{i=1}^{l} \frac{|\hat{f}_{n+i} - f_{n+i}|}{|\hat{f}_{n+i} + f_{n+i}|}$$

Более подробно используемый алгоритм приведен в работе Ю.И. Журавлева [2].

3 Вычислительный эксперимент

Условия измерения данных: данные — это измерения акселерометра и гироскопа, встроенных в мобильное устройство Redmi Note 5, хранящегося в переднем кармане куртки участника. Временные ряды содержат значения ускорения человека и углы ориентацию телефона для каждой из трёх осей — всего шесть временных рядов. Частота дискретизации составляет 200 Гц. Данные собраны одним участником, совершающим различные действия: ходьба и бег с различной скоростью, сидение за компьютером.

В эксперименте берется $l = 300$ и $n = T$.

В результате применения базового алгоритма $SMAPE(\mathbf{f}_{true}, \hat{\mathbf{f}}_{base}) = 0.144$,

а в модифицированном $SMAPE(\mathbf{f}_{true}, \hat{\mathbf{f}}_{mod}) = 0.199$.

Модификация метода нахождения ближайших соседей показала худшие результаты, чем безовый алгоритм. В любом случае на реальных данных оба метода работают не лучшим образом.

References

[1] McNames J.. 1999. Innovations in local modeling for time series prediction // *Ph.D. Thesis, Stanford University*

[2] Zhuravlev U.I, Ryazanov V. V., Senko O. V.. 2005. Recognition. Mathematical methods. Software system. Practical applications. // *Fazis, Moscow*

[3] N. P. Ivkin, M. P. Kuznetsov. 2015. Time series classification algorithm using combined feature description. . *Machine Learning and Data Analysis* (11):1471–1483.

[4] 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.

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

Received