Метрическое обучение в задаче многоклассовой классификации временных рядов

Исаченко Роман Владимирович

Научный руководитель: д. ф.-м. н. Стрижов В. В.

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

29 июня 2016 г.

Цель исследования

Задача

Определить вид активности человека по форме сигнала акселерометра мобильного телефона.

Требуется

Построить простой, устойчивый, точный алгоритм многоклассовой классификации временных рядов.

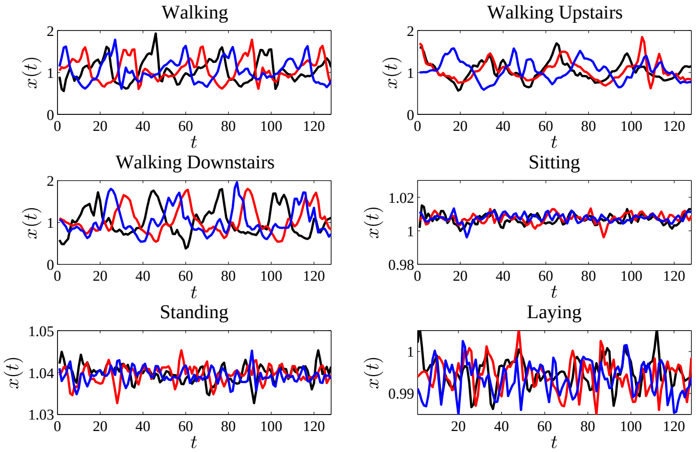
Предлагается

Для повышения качества классификации уточнить метрическое пространство временных рядов.

Метод

Обучение матрицы коварации множества временных рядов.

Временные ряды акселерометра



Литература

1. **Обзоры методов метрического обучения**
   Aurelien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. 2013.
   Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. 2006.
2. **Описание используемого алгоритма метрического обучения**
   K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. 2006.
3. **Описание алгоритма нахождения центроидов классов**
   Petitjean F. et al. Dynamic Time Warping averaging of time series allows faster and more accurate classification. 2014.

Постановка задачи

Дано: Выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ — множество объектов с известными метками классов. Каждый объект $\mathbf{x}_i \in \mathbf{X}$ —временной ряд, $\mathbf{X}$ — множество временных рядов фиксированной длины n, $y_i \in \{1, \dots, K\}$.

Требуется

Построить алгоритм $\mathbf{a} : \mathbf{x} \to \{1, \dots, K\}, \quad \forall \mathbf{x} \in \mathbf{X}$.

Модель имеет вид:

$$a = b \circ \mathbf{f} \circ G, \quad \text{где}$$

- G — процедура выравнивания временных рядов относительно центроида класса,
- $\mathbf{f}$ — алгоритм метрического обучения,
- b — алгоритм многоклассовой классификации.

Определение

Пусть $\mathbf{X}_e$ — множество временных рядов выборки $\mathfrak{D}$, принадлежащих одному классу e. Центроидом множества объектов $\mathbf{X}_e = \{\mathbf{x}_i | y_i = e\}_{i=1}^{\ell}$ по расстоянию ρ называется вектор:

$$\mathbf{c}_e = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^n} \sum_{\mathbf{x}_i \in \mathbf{X}_e} \rho(\mathbf{x}_i, \mathbf{c}).$$

$\rho(\cdot,\cdot)$ — стоимость выравнивающего пути между временными рядами.

Значение функции $\rho(\cdot,\cdot)$ вычисляется методом динамической трансформации шкалы времени.

Процедура выравнивания

Описание процедуры

- построить множество центроидов классов $\{\mathbf{c}_e\}_{e=1}^K$;
- по множеству центроидов найти пути наименьшей стоимости между каждым временным рядом $\mathbf{x}_i$ и центроидом его класса $\mathbf{c}_{y_i}$;
- по каждому пути восстановить выравненный временной ряд;
- привести множества выравненных временных рядов к нулевому среднему и нормировать на дисперсию.

Результатом является множество выравненных временных рядов.

Расстояние Махаланобиса

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j)},$$

где $\mathbf{A}$ — симметричная неотрицательно определённая матрица.

Представим матрицу $\mathbf{A}$ в виде:

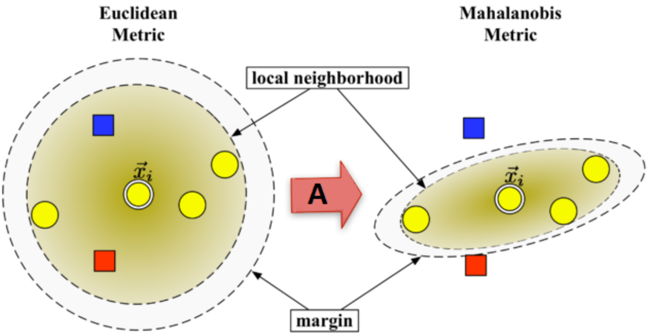
$$\mathbf{A} = \mathbf{L}^T \mathbf{L}.$$

Расстояние Махаланобиса - евклидово расстояние в новом пространстве объектов:

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j))^T (\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j))} = \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|_2.$$

Метрическое обучение состоит в нахождение данного пространства, т.е. матрицы $\mathbf{L}$.

Алгоритм LMNN (Weinberger, K. Q., 2006)



Объектом-нарушителем для $\mathbf{x}_i$ назовём объект $\mathbf{x}_l$ такой, что

$$\|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2 \leq \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 + 1, \quad \text{где } j \rightsquigarrow i,\ y_i \neq y_l.$$

$j \rightsquigarrow i$ означает, что $\mathbf{x}_j$ является одним из k ближайших соседей для $\mathbf{x}_i$.

Алгоритм LMNN

- Для каждого объекта минимизируем расстояния до k ближайших соседей из того же класса:

$$Q_1(\mathbf{L}) = \sum_{j \rightsquigarrow i} \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \to \min_{\mathbf{L}}.$$

- Штрафуем объекты-нарушители:

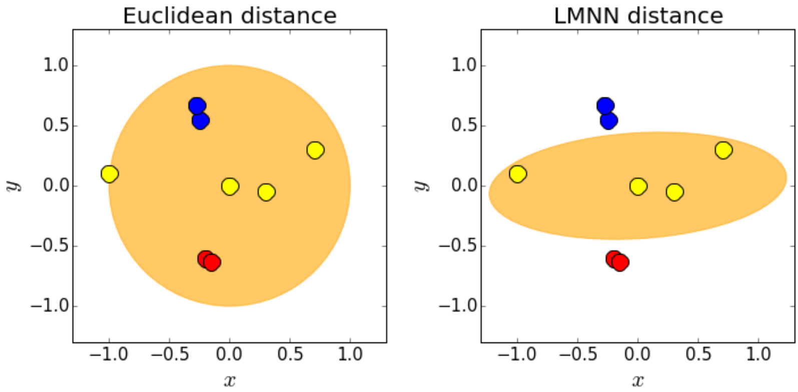
$$Q_2(\mathbf{L}) = \sum_{j \rightsquigarrow i} [y_i \neq y_l] \left[1 + \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2\right]_+ \to \min_{\mathbf{L}}.$$

Функция ошибки

$$Q(\mathbf{L}) = \mu Q_1(\mathbf{L}) + (1-\mu) Q_2(\mathbf{L}) \to \min_{\mathbf{L}},$$

где $\mu \in (0, 1)$ - весовой параметр.

Алгоритм LMNN



Классификация

- Пусть $\mathbf{x} \in \mathbf{X}$ — временной ряд без метки класса.
- Выравним временной ряд $\mathbf{x}$ относительно всех центроидов классов. Получим K временных рядов

$$\hat{\mathbf{x}}_e = G(\mathbf{x}, \mathbf{c}_e), \quad \text{где} \quad e = \{1, \dots, K\}.$$

- Отнесём временной ряд к классу, для которого минимально расстояние до соответствующего центроида.

Решающее правило

$$\hat{y} = \operatorname*{argmin}_{e \in \{1,\dots,K\}} d_{\mathbf{A}}(\hat{\mathbf{x}}_e, \mathbf{c}_e).$$

В качестве расстояния используем оптимальную метрику Махаланобиса с фиксированной матрицей $\mathbf{A}$.

Вычислительный эксперимент

Цель эксперимента

- синтетические данные: проверка работоспособности предложенного подхода;
- реальные данные: классификация вида активности человека по сигналу акселерометра мобильного телефона.

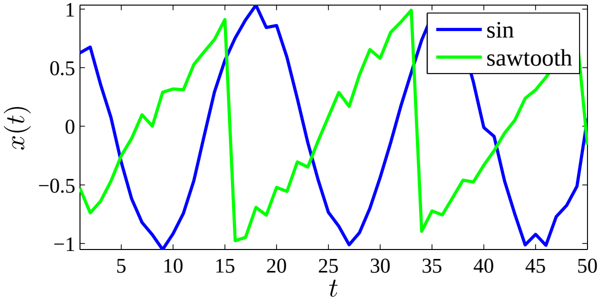
Синтетические временные ряды

- Количество классов: $K = 2$.
- Длина временного ряда: 40.
- Количество временных рядов в каждом классе: 50.

Описание классов

- 1 класс: $sin(kx + b)$, параметр b определяет сдвиг временного ряда.
- 2 класс: пилообразные функции.

Центроиды синтетических временных рядов



Точность классификации

- евклидова метрика: 73%
- метрика Махаланобиса: 94%

Вычислительный эксперимент

Реальные данные акселерометра

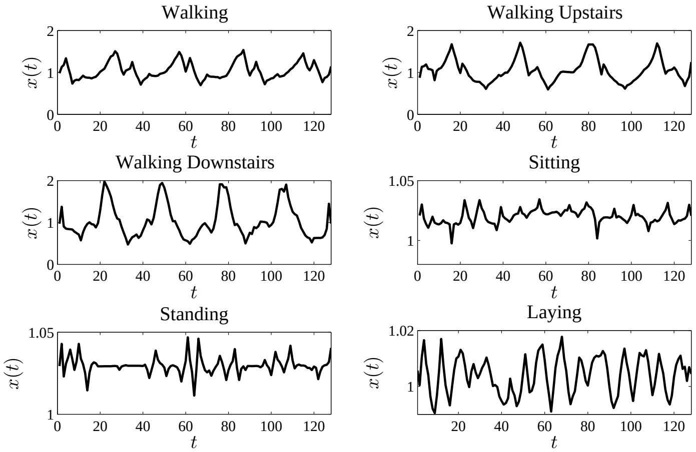
- Количество классов: $K = 6$.
- Длина временного ряда: 128.
- Количество временных рядов в каждом классе: 200.

Виды активности

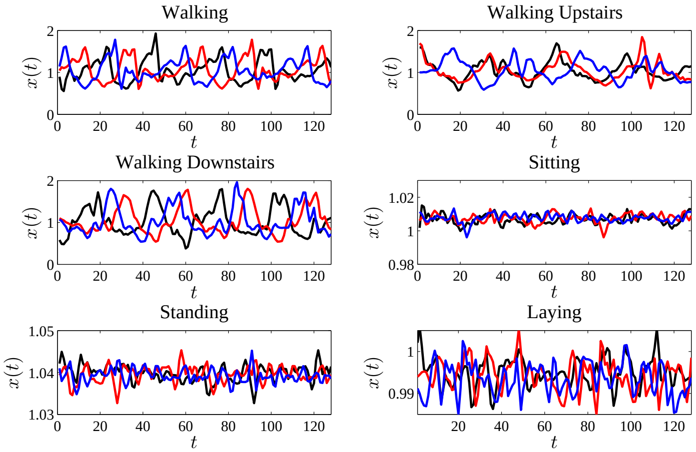
- ходьба;
- ходьба вверх;
- ходьба вниз;
- сидение;
- стояние;
- лежание.

https://archive.ics.uci.edu/ml/machine-learning-databases/00240/

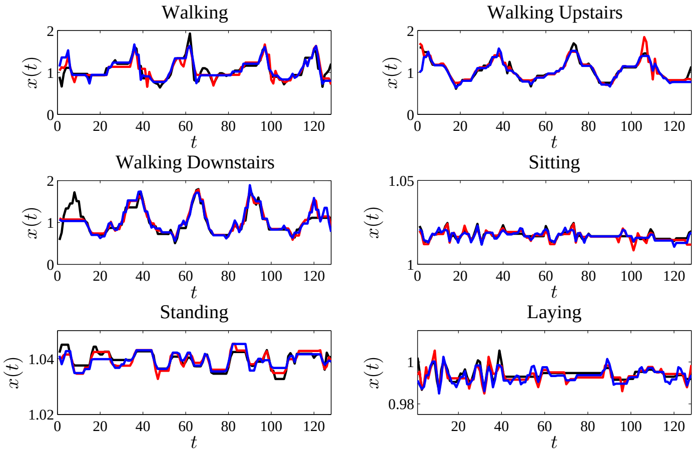
Центроиды временных рядов акселерометра



Временные ряды акселерометра



Выравненные временные ряды



Результаты эксперимента

Изменение качества классификации при использовании метрического обучения:

	Истинные метки классов					
	1	2	3	4	5	6
1	**0,355**	0,06	0,04	0	0	0
2	0,03	**0,43**	-0,095	0	0	0
3	0,02	0,025	**0,425**	0	0	0
4	0,015	-0,005	-0,025	**0,025**	0,025	0
5	-0,245	-0,25	-0,28	0	**-0,03**	-0,01
6	-0.175	-0,26	-0,06	-0,025	0,005	**-0,01**

Точность классификации

- евклидова метрика: 63.5%
- метрика Махаланобиса: 82.75%

Заключение

- Предложен новый подход к решению задачи многоклассовой классификации временных рядов.
- Проведён вычислительный эксперимент на синтетических данных.
- Проведён вычислительный эксперимент на реальных данных показаний акселерометра мобильного телефона.

Публикации

- Исаченко Р. В., Катруца А. М. Метрическое обучение в задачах кластеризации. // *JMLDA*, 2015.
- Исаченко Р. В., Стрижов В. В. Метрическое обучение в задаче многоклассовой классификации временных рядов. // *Информатика и её применения*, 2016.
- Motrenko A. P., Isachenko R. V., Neychev R. G. Feature generation for multiscale time series forecasting. // *ICDM*, 2016.