Классификация суперпозиций движений физической активности

Александр Денисович Проскурин
Евгений Александрович Белых

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 594, весна 2018

Цель исследования

Цель исследования

Найти способ распознать сложные движения человека, являющиеся суперпозицией более простых движений, используя данные акселерометра телефона.

Проблема

Данные являются непериодическими временными рядами, поэтому одной из задач является поиск способа описания временного ряда, а также способа выравнивания рядов друг относительно друга.

Новизна

Предлагается рассматривать движение как суперпозицию действий, а не однородный процесс.

Список литературы

- *Карасиков М.Е., Стрижов В.В.* Классификация временных рядов в пространстве параметров порождающих моделей // Информатика и ее применения, 2016
- *Кузнецов М.П., Ивкин Н.П.* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных, 2015
- *Фадеев И.В.* Выбор иерархических моделей в авторегрессионном прогнозировании // Магистерская диссертация, 2013, Московский физико-технический институт
- *Гончаров А.В.* Метрическая классификация временных рядов со взвешенным выравниванием относительно центроидов классов // 2015, Московский физико-технический институт

Пусть $D(X, Y)$ — это обучающая выборка, где (X, ρ) образует метрическое пространство временных рядов, Y — это метки временных рядов.
Пусть F — это функция построения множества признаков временного ряда:

$$F : X \to \mathbb{R}^n$$

Пусть G — многоклассовый классификатор, который переводит признаки в метки:

$$G : \mathbb{R}^n \to Y$$

Итоговый алгоритм — это композиция некоторого G и F:

$$a = G \circ F$$

Функционал качества

Пусть задана некоторая функция потерь $L : X \times Y \times Y \to \mathbb{R}$, тогда функционал качества имеет вид:

$$Q(a, D) = \frac{1}{|D|} \sum_{(x,y) \in D} L(x, a(x), y).$$

Предлагается в качестве функции потерь использовать индикатор:

$$Q(a, D) = \frac{1}{|D|} \sum_{(x,y) \in D} I(a(x) \neq y)$$

В методе обучения будем сначала фиксировать F, а после оптимизировать функцию G, тогда наш оптимальный алгоритм принимает следующий вид $a_F = \widehat{G} \circ F$, где

$$\widehat{G} = \underset{G}{\operatorname{argmin}}(Q(G \circ F, D))$$

Оценка эффективности

Чтобы оценить эффективность метода обучения, мы будем разбивать нашу выборку r раз на тестовую и тренировочную ($D = A_1 \cup B_1 = A_2 \cup B_2 = ... = A_r \cup B_r$), в таком случае наш критерий качества a_F будет:

$$QV(a_F, D) = \frac{1}{r} \sum_{i=1}^{r} Q(a_F(A_i), B_i)$$

Итоговая цель — найти

$$\widehat{a} = \underset{F}{\operatorname{argmin}}(QV(a_F, D))$$

Базовый алгоритм

Для начала проведем базовый вычислительный эксперимент. Используем несколько известных методов, их модификации и готовый датасет для тестирования.

Рассмотрим параметрическую модель, которая будет приближать реальные значения нашего временного ряда:

$$g(w, X) \to X, \text{ где } w \in \mathbb{R}^n.$$

В качестве параметрической модели рассмотрим:

- Авторегрессионную модель $AR(p)$:

  Пусть $x = [x_1, x_2, ...x_t]$ — временной ряд, где $x_i \in \mathbb{R}$ Тогда
  $$g(w, x) = [\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_t],$$
  $$\text{где } \widehat{x}_k = \begin{cases} x_k, & k = 1, ..., p, \\ w_0 + \sum_{i=1}^{p} w_i \cdot x_{k-i}, & k = p+1, ..., t. \end{cases}$$
- Преобразование Фурье

Алгоритм классификации

Для завершения построения алгоритма классификации временных рядов, необходимо построить классификатор G по обучающей выборке $\{(F(x), y) \mid (x, y) \in D\}$.

В качестве классификатора используем следующие модели:

- KNN
- Random forest
- Logistic regression

Базовый вычислительный эксперимент

В базовом вычислительном эксперименте предполагается использовать следующие данные:

- Датасет WISDM для базового вычислительного эксперимента

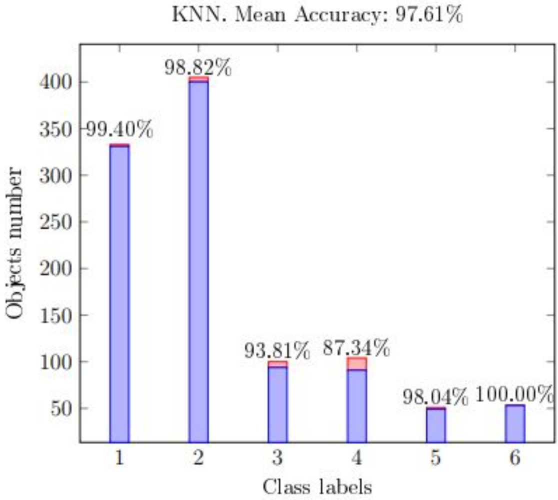
Датасет WISDM

Датасет WISDM содержит показания акселерометра для шести видов человеческой активности:

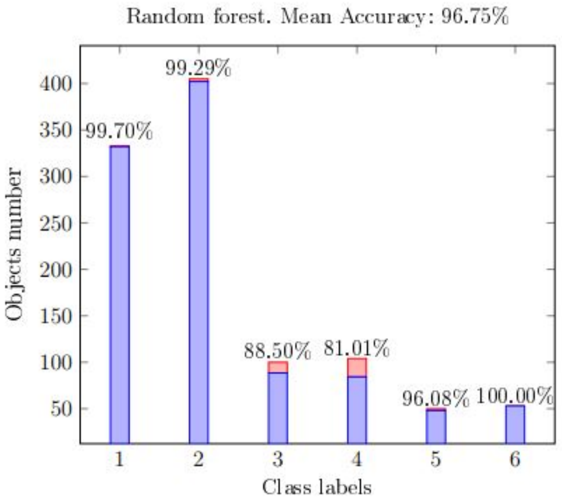
1. Jogging
2. Walking
3. Upstairs
4. Downstairs
5. Sitting
6. Standing

Необработанные данные, представляющие из себя последовательность размеченных показаний акселерометра, были разбиты на временные ряды длиной по 200 отсчетов (10 секунд).

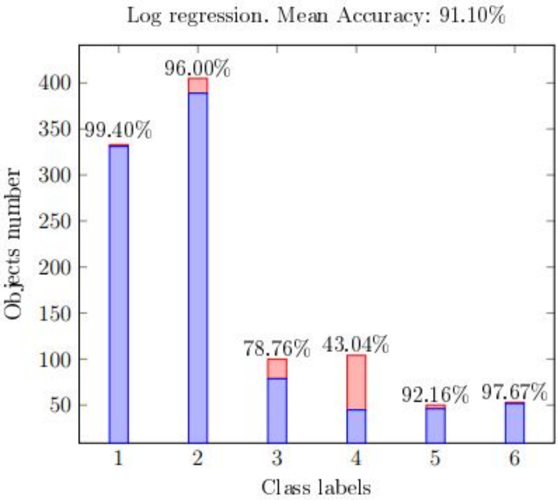
Результаты при использовании классификатора KNN:



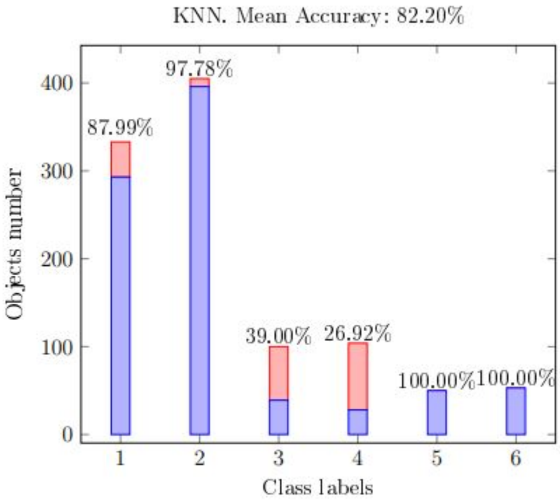
Результаты при использовании классификатора Random forest:



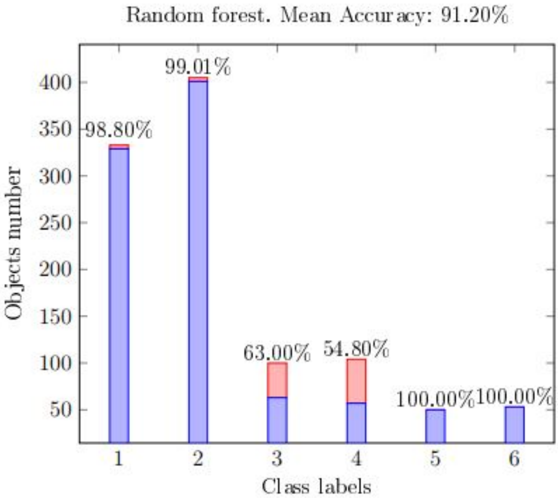
Результаты при использовании классификатора Logistic regression:



Результаты при использовании классификатора KNN:

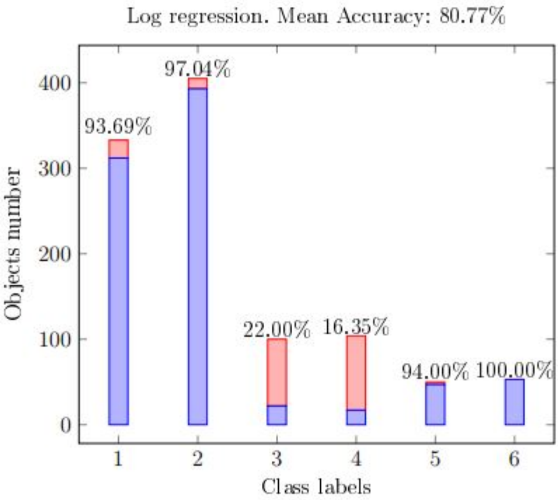


Результаты при использовании классификатора Random forest:



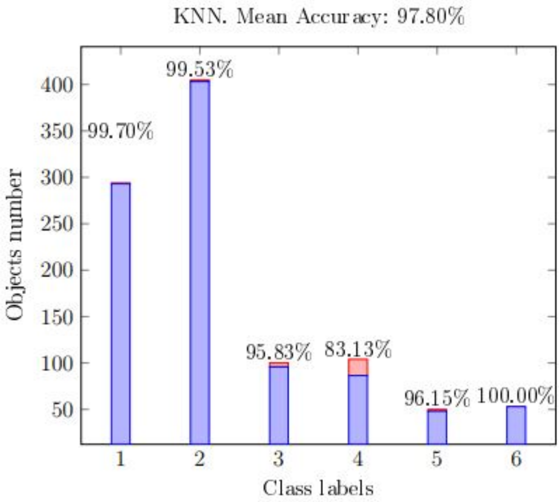
Преобразование Фурье

Результаты при использовании классификатора Logistic regression:



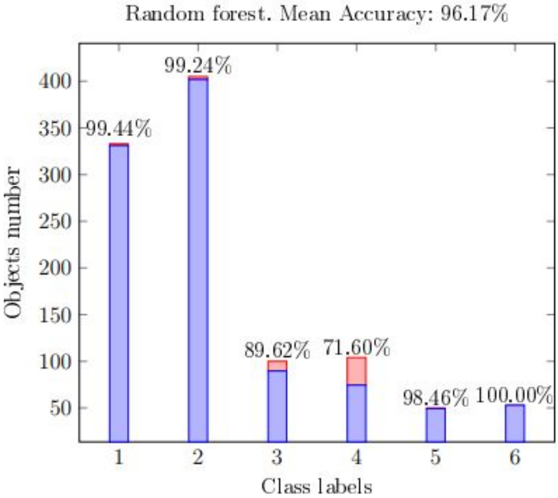
Комбинирование линейного преобразования рядов и модели авторегрессии

Результаты при использовании классификатора KNN:



Комбинирование линейного преобразования рядов и модели авторегрессии

Результаты при использовании классификатора Random forest:



Итоги базового эксперимента

Как видно, все методы показывают хорошие результаты. В то же время, наилучшие результаты показывает модель комбинирования преобразования рядов и авторегрессии. При этом, среди классификаторов наилучшие результаты показывает KNN. Также можно заметить, что наихудшие результаты алгоритмы показывают для классов «Upstairs» и «Downstairs». Кроме того, все алгоритмы можно улучшать с помощью, например, добавления дополнительных признаков или перебора параметров классификаторов.

Модель **SEMOR** решает задачу аппроксимации отнормированного временного ряда Z временным рядом $\mathbf{x}$. Для этого производится сдвиг и растяжение ряда $\mathbf{x}$:

$$\hat{\mathbf{x}}(t) = \mathbf{x}(\omega_1 * t + \omega_2).$$

Задача заключается в поиске таких коэффициентов ω_1, ω_2, ω_3 и ω_4, что величина

$$||Z - (\omega_3 * \mathbf{x}(\omega_1 * t + \omega_2) + \omega_4)||$$

минимизируется.

Модель локальной аппроксимации SEMOR для двух рядов

Пусть Z — отнормированный ряд, который необходимо классифицировать.
$\{G_1, G_2, ..., G_n\}$ и $\{H_1, H_2, ..., H_m\}$ — два множества отнормированных рядов.
Ряд Z нужно аппроксимировать суперпозицией некотой пары рядов G_i и H_j.
Суперпозицией преобразованных рядов является их сумма, которая также преобразована с помощью сдвига и растяжения:

$$Q = \omega_5 * (G_i(\omega_1 * t + \omega_2) + H_j(\omega_3 * t + \omega_4)) + \omega_6$$

Таким образом, необходимо найти:

$$argmin_{(i,j)}||Q - Z||,$$

где все ω в Q выбраны так, что они минимизируют значение функции.

Путь наименьшей стоимости между рядами DTW

Пусть G и H — временные ряды из $\mathbb{R}^n$ и $\mathbb{R}^m$.
Пусть Ω^{nm} — матрица, такая что $\Omega_{ij} = (G_i - H_j)^2$.
Путем будем назвать последовательность пар индексов Ω:
$\pi = \{\pi_r\} = \{(i_r, j_r)\}$.
Стоимость пути π: $Cost(G, H, \Pi) = \sum_{(i,j)\in\pi} \Omega_{ij}$.
Тогда задача поиска кратчашего пути заключается в поиске пути:

$$\hat{\pi} = argmin_{\pi} Cost(G, H, \pi)$$

а точнее его величины $\rho(G, H) = Cost(G, H, \hat{\pi})$.

Построим новую матрицу Γ^{nm} следующим образом:

$$\Gamma_{1j} = \Omega_{1j},\ \Gamma_{i1} = \Omega_{i1},\ i \in \{1, .., n\},\ j \in \{1, .., m\}.$$
$$\Gamma_{ij} = \Omega_{ij} + min(\Gamma_{i-1,j}, \Gamma_{i,j-1}, \Gamma_{i-1,j-1})$$

Функцией расстояния между рядами G и H будем считать стоимость пути между ними:

$$\rho(G, H) = \Gamma_{nm}$$

.

Путь наименьшей стоимости между рядами DTW для двух рядов

Пусть есть ряды G из $\mathbb{R}^n$, H из $\mathbb{R}^m$, Z из $\mathbb{R}^t$.
Пусть, как и в случае с двумя рядами, есть матрица расстояний Ω^{nmt}, где $\Omega_{ijk} = (G_i - H_j - Z_k)^2$.
Определим путь как последовательность троек индексов Ω: $\pi = \{\pi_r\} = \{(i_r, j_r, k_r)\}$.
Стоимость пути π: $Cost(G, H, Z, \pi) = \sum_{(i,j,k)\in\pi} \Omega_{ijk}$.
Тогда задача заключается в поиске пути

$$\hat{\pi} = argmin_{\pi} Cost(G, H, Z, \pi)$$

а точнее его величины

$$\rho(G, H, Z) = Cost(G, H, Z, \hat{\pi})$$

.

Путь наименьшей стоимости между рядами DTW для двух рядов

Построим новую матрицу Γ^{nmt}, элементы которой определим следующим образом:

$$\Gamma_{1j1} = \Omega_{1j1}, \Gamma_{i11} = \Omega_{i11}, \Gamma_{11k} = \Omega_{11k}$$

$$i \in \{1, .., n\}, j \in \{1, .., m\}, k \in \{1, .., t\}$$

$$\Gamma_{ijk} = \Omega_{ijk} + min(\Gamma_{i-1,j,k}, \Gamma_{i,j-1,k}, \Gamma_{i,j,k-1}, \Gamma_{i-1,j-1,k}, \Gamma_{i-1,j,k-1}, \Gamma_{i,j-1,k-1}, \Gamma_{i-1,j-1,k-1})$$

В таком случае в качестве значения функции расстояния между рядами G, H и Z будем считать стоимость пути между ними:

$$\rho(G, H, Z) = \Gamma_{nmt}.$$

Итоговый алгоритм

Пусть Z — ортонормированный ряд, который необходимо классифицировать.
Пусть $\{G_1, G_2, ..., G_n\}$ и $\{H_1, H_2, ..., H_m\}$ — два множества отнормированных рядов, причем Z является суперпозицией некоторой пары рядов G_i и H_j.
Будем перебирать пары рядов из множеств $\{G_i\}$ и $\{H_j\}$.
Для каждой фиксированной пары (G_i, H_j) и ряда Z построим матрицу Г методом **DTW**.
Пусть кратчайший путь по матрице Г — это $\hat{\pi}$.
Составим новые ряды $\hat{G}$, $\hat{H}$, $\hat{Z}$, где $\hat{G} = \{G_i\}$ по $i \in \hat{\pi}$, $\hat{H} = \{H_j\}$ по $j \in \hat{\pi}$, $\hat{Z} = \{Z_k\}$ по $k \in \hat{\pi}$.

Итоговый алгоритм

С помощью линейной регрессии построим:

$$\hat{Q} = \omega_5 * (\hat{G} + \hat{H}) + \omega_6$$

В конечном итоге нужно выбрать такую пару G_i и H_j, для которой будет минимальным значение величины:

$$||\hat{Q} - Z||$$

Для проверки качества построенного алгоритма, протестируем его на искусственных данных.
Сгенерируем два случайных ряда A и B. Возьмем множества рядов:

- $A_1 = A/8$
- $A_2 = sin(A)$
- $A_3 = log(A)$
- $A_4 = exp(A/10)$
- $B_1 = cos(B)$
- $B_2 = sin(log(B))$
- $B_3 = sin(B * 2)$
- $B_4 = 2 * tg(B/5)$

В качестве рядов для классификации Z будем рассматривать суммы рядов A_i и B_j. Сгенерировав эти суммы, запустим построенный алгоритм на них.

Вычислительный эксперимент

Качество работы итогового алгоритма на искусственных данных:

A \ **B**	B_1	B_2	B_3	B_4
A_1	100%	100%	90%	60%
A_2	100%	80%	100%	40%
A_3	90%	90%	80%	60%
A_4	80%	100%	50%	100%

В среднем качество работы: 82.5%

В работе была рассмотрена задача классификации видов физической активности человека. Был проведен вычислительный эксперимент с базовыми алгоритмами и их модификациями, который показал очень хорошие результаты. Был также предложен метод классификации суперпозиции временных рядов с помощью комбинации модели **SEMOR** и алгоритма **DTW**.
Построенный алгоритм показал хорошие результаты на искусственных данных. Ожидается, что такие же хорошие результаты будут показаны и на реальных временных рядах, описывающих физическую активность человека, что в перспективе может привести к построению алгоритма лучшего, чем базовый.