

Abstract

This paper discusses a problem of time series classification in case of several classes. The proposed classification model uses the matrix of distance between time series. This distance measure is defined by dynamic time warping method. The dimension of the distance matrix is very high. This paper introduces centroids of each class as a reference objects to decrease this dimension. The distance matrix with lower dimension describes the distance between all objects and reference objects. We use this method for human activity recognition and investigate the quality of classification on data from the mobile accelerometer. This metric algorithm of classification is compared with separating classification algorithm.

Keywords: *metric classification, dynamic time warping, time series classification, centroid, distance function.*

¹1 — Moscow Institute of Physics and Technology, alex.goncharov@phystech.edu
2 — Moscow Institute of Physics and Technology, maria_popova@phystech.edu
3 — Dorodnicyn Computing Centre of the Russian Academy of Sciences, strijov@ccas.com

Metric time series classification using dynamic warping relative to centroids of classes

A. Goncharov¹, M. Popova², V. Strijov³

July 15, 2015

References

- [1] *Kwapisz J. R., Weiss G. M., Moore S.* Activity recognition using cell phone accelerometers // SIGKDD Explorations, 2010. Vol. 12, No. 2. Pp. 74–82.
- [2] *Kwapisz J. R.* Data from accelerometer// Sourceforge.net. 2010. http://sourceforge.net/p/mlalgorithms/TSLearning/data/preprocessed_large.csv
- [3] *Hoseini-Tabatabaei S. A., Gluhak A., Tafazolli R.* A survey on smartphone-based systems for opportunistic user context recognition // ACM Computing Surveys (CSUR), 2013. Vol. 45, No. 3, P. 27.
- [4] *Kuznecov M. P.* Priznakovaja klassifikacii vremennyh rjadov// Sourceforge.net. 2014. <https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/TSLearning/doc/TSClassification/TSClassification.pdf>
- [5] *Popova M. S., Strizhov V. V.* Vybór optimal'noj modeli klassifikacii fizicheskoj aktivnosti po izmerenijam akselerometra // Informatika i ee primeneniya, 2015. T. 9. Vol 1. P. 79–89.
- [6] *Faloutsos C., Ranganathan M., Manolopoulos Y.* Fast Subsequence Matching in Time-Series Databases // ACM International Conference on Management of Data (SIGMOD), 1994. Pp. 419-429.
- [7] *Berndt D. J., Clifford J.* Using dynamic time warping to find patterns in time series // Workshop on Knowledge Discovery in Databases (KDD Workshop), 1994. Pp. 359-370.
- [8] *Keogh E. J., Ratanamahatana C. A.* Exact indexing of dynamic time warping. // Knowl. Inf. Syst., 2005. Vol. 7, No. 3, Pp. 358-386.
- [9] *Vlachos M., Gunopulos D., Kollios G.* Discovering similar multidimensional trajectories // IEEE International Conference on Data Engineering (ICDE), 2002. Pp. 673-684.

- [10] *Chen L., Ng R. T.* On the marriage of lp-norms and edit distance // Very Large Data Bases (VLDB), 2004. Pp. 792-803
- [11] *Chen L., Özsu M. T. Oria V.* Robust and fast similarity search for moving object trajectories // ACM International Conference on Management of Data (SIGMOD), 2005. Pp. 491-502.
- [12] *Frentzos E., Gratsias K., Theodoridis Y.* Index-based most similar trajectory search // IEEE International Conference on Data Engineering (ICDE), 2007. Pp. 816-825.
- [13] *Morse M. D., Patel J. M.* An efficient and accurate method for evaluating time series similarity // ACM International Conference on Management of Data (SIGMOD), 2007. Pp. 569-580.
- [14] *Chen Y., Nascimento M. A., Ooi B. C., Tung A. K. H.* SpADe: On Shape-based Pattern Detection in Streaming Time Series // IEEE International Conference on Data Engineering (ICDE), 2007. Pp. 786-795
- [15] *Scheuermann P., Wang X., Ding H., Trajcevski G., Keogh E.* Querying and mining of time series data: experimental comparison of representations and distance measures // Very Large Data Bases (VLDB), 2008. Pp. 1542-1552.
- [16] *Keogh E. J., Pazzani M. J.* Scaling up Dynamic Time Warping to Massive Datasets // Principles of Data Mining and Knowledge Discovery, 1999. Pp. 1-11.
- [17] *Salvador S., Chan P.* Fastdtw: Toward accurate dynamic time warping in linear time and space // Workshop on Mining Temporal and Sequential Data, 2004. P. 11.
- [18] *Goncharov A. V.* Programm realization // Sourceforge.net. 2015. <http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group274/Goncharov2015MetricClassification/code/main.m>

Метрическая классификация временных рядов с выравниванием относительно центроидов классов*

А. В. Гончаров¹, М. С. Попова², В. В. Стрижов³

Аннотация: В работе рассматривается задача многоклассовой классификации временных рядов. Классификация производится с помощью метрических методов, использующих матрицу попарных расстояний между временными рядами. Вычисление такой матрицы является трудоемким, так как ее размерность равна числу объектов выборки. С целью снижения размерности предлагается предварительно выделять эталонные объекты, а именно центроиды каждого класса, и строить матрицу попарных расстояний между объектами выборки и эталонными объектами классов. Для вычисления попарных расстояний между объектами предлагается использовать метод динамического выравнивания временных рядов. В качестве прикладной задачи рассматривается задача распознавания типа движения по данным акселерометра мобильного телефона. Метрический алгоритм классификации, исследованный в этой работе, сравнивается в точности с алгоритмом разделяющей классификации.

Ключевые слова: метрическая классификация, динамическое выравнивание, классификация временных рядов, центроид, функция расстояния.

Введение

В работе решается задача идентификации движений человека по временным рядам акселерометра [1]. Данные [2] представляют собой измерения акселерометра некоторого носимого устройства, например мобильного телефона, находящегося в кармане человека, и используются для идентификации действия человека в конкретный момент времени [3]. Метки классов движений служат: ходьба, бег, подъем по лестнице, спуск по лестнице, сидение, лежание.

*Работа выполнена при финансовой поддержке РФФИ, проект 13-07-00709, и в рамках инициативы "МИТ-Сколтех".

¹Московский физико-технический институт, alex.goncharov@phystech.edu

²Московский физико-технический институт, maria_popova@phystech.edu

³Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра "Информатика и управление" Российской академии наук, strijov@ccas.ru

В работах [4, 5] приведены методы, решающие задачу классификации временных рядов акселерометра. В [4] исследованы различные методы построения признакового описания временных рядов, в частности, метод экспертного построения признаков и метод построения признакового описания на основе гипотезы порождения данных. Результаты показывают, что построенное признаковое пространство достаточным образом описывает зависимую переменную и приводит к высокой точности классификации. В [5] для решения задачи использованы нейронные сети с небольшим числом связей между нейронами, которые достаточно точно решают задачу классификации и обладают свойством устойчивости к возмущениям данных.

Особенность этой работы заключается в использовании метрических способов классификации с предварительным отбором эталонных объектов. Предлагается ввести функции расстояния между объектами выборки, определить признаковое пространство при помощи этой функции и провести дальнейшую классификацию объектов методом k ближайших соседей.

Функция расстояния между временными рядами может быть задана различными способами: Евклидово расстояние [6], метод динамического выравнивания временных рядов [7, 8], метод, основанный на нахождении наибольшей общей последовательности [9], Edit Distance with Real Penalty [10], Edit Distance on Real sequence [11], DISSIM [12], Sequence Weighted Alignment model [13], Spatial Assembling Distance [14] и другие.

В [15] показано, что разность между значениями временного ряда, соответствующими различным временным отсчетам, не может рассматриваться в качестве описания расстояния между двумя объектами: эта мера расстояния чувствительна к шуму и локальным временным сдвигам. Для решения задачи предлагается использовать метод динамического выравнивания временных рядов (англ. Dynamic Time Warping) [16]. Как показано в [17], этот метод находит наилучшее соответствие между двумя временными рядами, если они нелинейно деформированы друг относительно друга — растянуты, сжаты или смещены вдоль оси времени. Метод DTW используется для определения сходства между ними и введения расстояния между двумя объектами.

Для снижения размерности признакового пространства предлагается искать расстояния от некоторого объекта до эталонных объектов каждого класса. Этими эталонными объектами предлагается назначать центроиды классов. Для определения центроида требуется одновременно выровнять все временные ряды, принадлежащие данному классу. Эта задача обладает степенной сложностью — t^2n^2 , где t — число объектов класса, n — длина каждого временного ряда.

Процедура классификации выполняется в три шага. Первый — отбор эталонных объектов каждого класса. Второй — построение матрицы попарных расстояний сниженной размерности между временными рядами и эталонными объектами каждого класса. Третий — классификация временных рядов методом k ближайших соседей с помощью матрицы попарных расстояний.

С целью сравнения результатов классификации методом, предложенным в этой работе, и результатов, полученных в [5], проведен вычислительный эксперимент на

исходных данных [2].

1 Постановка задачи

Задана выборка $\mathcal{D} = \{(\mathbf{s}_i, y_i)\}_{i=1}^m$, состоящая из пар объект-ответ. Объектами являются временные ряды $\mathbf{s}_i \in \mathbb{R}^n$, а ответами являются метки класса — $y_i \in Y = \{1, \dots, E\}$, где $E \ll m$. Примеры временных рядов показаны на рисунке 1. Выборка разбита на обучающую \mathcal{D}_l и контрольную \mathcal{D}_t .

Принята функция ошибки S . Требуется построить модель классификации $f : \mathbb{R}^n \rightarrow Y$, минимизирующую функцию ошибки S на контрольной выборке:

$$f = \underset{f}{\operatorname{argmin}}(S(f(\mathbf{s}_i), y_i | \mathcal{D}_t)) = \underset{f}{\operatorname{argmin}} \left(\frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{|\mathcal{D}_t|} [f(\mathbf{s}_i) \neq y_i] \right). \quad (1)$$

2 Описание алгоритма метрической классификации

Модель метрической классификации состоит из трех функций. Первая функция выделяет центроиды классов, вторая функция строит матрицу попарных расстояний между центроидами и объектами, а третья функция классифицирует объекты по полученной матрице.

2.1 Отбор эталонных объектов

В общем случае матрица попарных расстояний \mathbf{A} имеет размерность $m \times m$, где m — число объектов выборки \mathcal{D} . Для снижения размерности матрицы попарных расстояний предлагается строить ее не для всех пар элементов выборки \mathcal{D} , а для пар $(\mathbf{s}_i, \mathbf{z}_e)$, где \mathbf{s}_i — элемент выборки, а \mathbf{z}_e — эталонный объект класса e из Y . Предлагается в качестве эталонного объекта использовать *центроид* этого класса. Это решение даст значительное снижение размерности матрицы попарных расстояний до $m \times E$, где E — число меток классов.

Определение 1 Пусть \mathcal{D}_e — множество элементов из \mathcal{D} , принадлежащих одному классу e из Y . Центроидом множества векторов $\mathcal{D}_e = \{\mathbf{s}_i | y_i = e\}_{i=1}^m$ по расстоянию ρ назовем вектор $\mathbf{z}_e \in \mathbb{R}^n$ такой, что:

$$\mathbf{z}_e = \underset{\mathbf{z} \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i \in \mathcal{D}_e} \rho(\mathbf{s}_i, \mathbf{z}).$$

Опишем метод оценки *центроида* \mathbf{z}_e из множества векторов $\mathcal{D}_e \subset \mathbb{R}^n$. Строим матрицу попарных расстояний \mathbf{A} между всеми элементами множества \mathcal{D}_e . С помощью метода главных компонент снижаем размерность пространства столбцов матрицы \mathbf{A} .

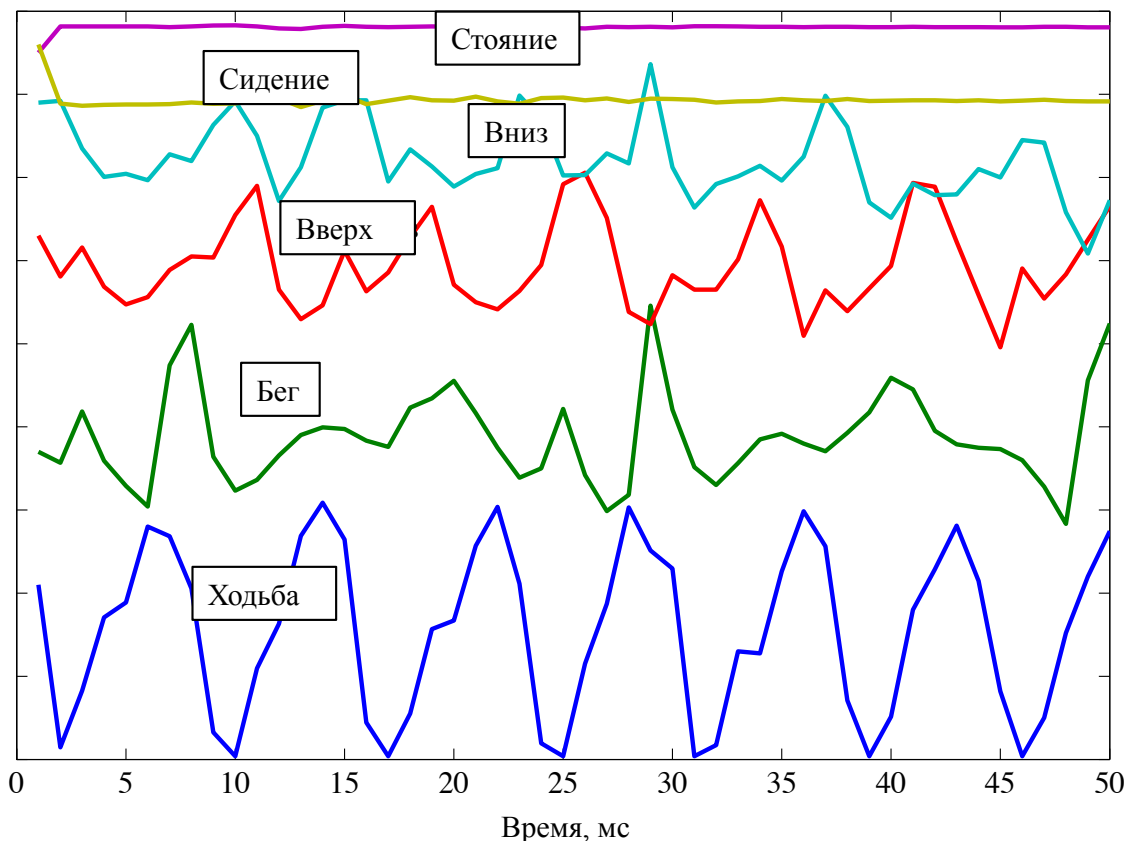


Рис. 1: Примеры временных рядов, измерений акселерометра для разных видов физической активности

Построим такие проекции исходных объектов на новую систему координат, называемых главными компонентами, чтобы сумма квадратов расстояний от объектов до их проекций на первые h ($h < n$) главных компонент была минимальной. Опишем этот метод более подробно. Рассмотрим такую ортогональную матрицу \mathbf{W} в линейной комбинации $\mathbf{B} = \mathbf{W}\mathbf{A}^T$ векторов-строк матрицы \mathbf{A} , что векторы-столбцы $\mathbf{b}_1, \dots, \mathbf{b}_n$ матрицы \mathbf{B} будут иметь максимальную сумму их дисперсий:

$$\sum_{j=1}^n \sigma^2(\mathbf{b}_j) \rightarrow \max, \quad \text{где} \quad \sigma^2(\mathbf{b}) = \frac{1}{m} \sum_{i=1}^m (\mathbf{b}_i - \bar{\mathbf{b}})^2, \quad \bar{\mathbf{b}} = \frac{1}{m} \sum_{i=1}^m \mathbf{b}_i.$$

Матрица попарных расстояний между объектами в пространстве меньшей размерности $\mathbf{A}^T = \mathbf{W}'\mathbf{A}^T$ есть проекция векторов-строк матрицы \mathbf{A} на h первых главных компонент.

Центроидом множества векторов \mathfrak{D}_e по расстоянию ρ будем приближенно считать вектор $\mathbf{z}_e \in \mathfrak{D}_e$ такой, что в пространстве первых h главных компонент \mathfrak{D}_e он является ближайшим к остальным объектам, то есть:

$$\mathbf{z}_e = \operatorname{argmin}_{\mathbf{z} \in \mathfrak{D}_e} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \rho(\mathbf{W}'^T \mathbf{s}_i, \mathbf{W}'^T \mathbf{z}). \quad (2)$$

Рассматривая подмножества объектов каждого класса по отдельности, вводим функцию g_ρ , которая каждому классу из Y ставит в соответствие центроид данного класса по расстоянию ρ по формуле (2)

$$g_\rho : Y \rightarrow \mathfrak{D}',$$

где $\mathfrak{D}' \subset \mathfrak{D}_l$ — множество центроидов каждого класса.

2.2 Введение функции расстояния между объектами

В данной работе в качестве метрического расстояния между объектами предлагается использовать стоимость *пути наименьшей стоимости* между этими объектами.

Дадим определение пути наименьшей стоимости. Пусть $\Omega^{n \times n}$ — это матрица, такая что ее элемент Ω_{ij} равен разности между i -ым и j -ым элементами последовательностей \mathbf{s}_1 и \mathbf{s}_2 :

$$\Omega_{ij} = |\mathbf{s}_{1i} - \mathbf{s}_{2j}|.$$

Определение 2 *Путем π между последовательностями \mathbf{s}_1 и \mathbf{s}_2 назовем упорядоченное множество пар индексов элементов матрицы Ω :*

$$\pi = \{\pi_r\} = \{(i_r, j_r)\}, \quad r = 1, \dots, R, \quad i, j \in \{1, \dots, n\},$$

где R — длина пути, зависящая от выбора пути. Он должен удовлетворять следующим условиям:

Граничные условия. $\pi_1 = (1, 1)$ и $\pi_R = (n, n)$, т. е. начало и конец π находятся на диагонали в противоположных углах Ω .

Непрерывность. Пусть $\pi_r = (p_1, p_2)$ и $\pi_{r-1} = (q_1, q_2)$, $r = 2, \dots, R$. Тогда

$$p_1 - q_1 \leq 1, \quad p_2 - q_2 \leq 1.$$

Это ограничение нужно, чтобы в шаге пути π участвовали только соседние элементы матрицы (включая соседние по диагонали).

Монотонность. Пусть $\pi_r = (p_1, p_2)$ и $\pi_{r-1} = (q_1, q_2)$, $r = 2, \dots, R$. Тогда выполняется хотя бы одно из условий:

$$p_1 - q_1 \geq 1, \quad p_2 - q_2 \geq 1.$$

Это ограничение обусловлено природой рассматриваемых последовательностей и предназначено для монотонности функции выравнивания времени.

Определение 3 *Стоимостью $Cost(\mathbf{s}_1, \mathbf{s}_2, \pi)$ пути π между последовательностями \mathbf{s}_1 и \mathbf{s}_2 назовем*

$$Cost(\mathbf{s}_1, \mathbf{s}_2, \pi) = \sum_{(i,j) \in \pi} \Omega_{ij}.$$

Определение 4 *Путем наименьшей стоимости (выравнивающим путем) $\hat{\pi}$ между последовательностями \mathbf{s}_1 и \mathbf{s}_2 назовем путь, который имеет наименьшую стоимость среди всех возможных путей между последовательностями \mathbf{s}_1 и \mathbf{s}_2 :*

$$\hat{\pi} = \underset{\pi}{\operatorname{argmin}} Cost(\mathbf{s}_1, \mathbf{s}_2, \pi).$$

Обозначим стоимость выравнивающего пути между последовательностями \mathbf{s}_1 и \mathbf{s}_2 через $\delta(\mathbf{s}_1, \mathbf{s}_2) = Cost(\mathbf{s}_1, \mathbf{s}_2, \hat{\pi})$.

Для вычисления стоимости выравнивающего пути в данной работе используется метод DTW. Согласно этому методу необходимо построить новую матрицу γ , элементы которой определяются следующим образом:

$$\gamma_{1j} = \Omega_{1j}, \quad \gamma_{i1} = \Omega_{i1}, \quad i, j = 1, \dots, n,$$

$$\gamma_{ij} = \Omega_{ij} + \min(\gamma_{i,j-1}, \gamma_{i-1,j}, \gamma_{i-1,j-1}), \quad i, j = 2, \dots, n.$$

Элемент γ_{ij} матрицы γ равен стоимости выравнивающего пути между последовательностями $\{s_{1a}\}_{a=1}^i$ и $\{s_{2a}\}_{a=1}^j$.

В качестве значения функции расстояния между двумя объектами выберем стоимость выравнивающего пути между ними из Определения 3:

$$\rho(\mathbf{s}_1, \mathbf{s}_2) = \delta(\mathbf{s}_1, \mathbf{s}_2) = \gamma_{nn}.$$

Иначе, для множеств декартова произведения $\mathfrak{D} \times \mathfrak{D}'$:

$$\rho : \mathfrak{D} \times \mathfrak{D}' \rightarrow \mathbf{X},$$

где \mathbf{X} — матрица попарных расстояний между объектами и центроидами классов. Ее элементы равны $X_{ij} = \rho(\mathbf{s}_i, \mathbf{d}_j)$, $\mathbf{s}_i \in \mathfrak{D}$, $\mathbf{d}_j \in \mathfrak{D}'$.

2.3 Описание алгоритма классификации

Считаем матрицу \mathbf{X} множеством вектор-строк, являющихся описанием объектов $\{\mathbf{x}_i\}_{i=1}^m$. Матрица \mathbf{X} используется как матрица объект-признак в алгоритме классификации. Применяется метод k ближайших соседей. При этом, $\mathbf{X}^l = \rho(\mathfrak{D}^l \times \mathfrak{D}')$ — обучающая выборка из множества \mathbf{X} , а $\mathbf{X}^t = \rho((\mathfrak{D} \setminus \mathfrak{D}^l) \times \mathfrak{D}')$ — тестовая. Каждому объекту обучающей выборки $\mathbf{x}_i \in \mathbf{X}^l$, соответствующему исходному объекту $\mathbf{s}_i \in \mathfrak{D}$, поставлена в соответствие метка класса y_i .

Для произвольного объекта $\mathbf{u} \in \mathbf{X}$ расположим элементы обучающей выборки $\mathbf{x}_1, \dots, \mathbf{x}_{|X^l|} = X^l$ в порядке возрастания расстояний до \mathbf{u} :

$$\eta(\mathbf{u}, \mathbf{x}_{\mathbf{u}}^{(1)}) \leq \eta(\mathbf{u}, \mathbf{x}_{\mathbf{u}}^{(2)}) \leq \dots \leq \eta(\mathbf{u}, \mathbf{x}_{\mathbf{u}}^{(|X^l|)}),$$

где $\eta(\mathbf{u}, \mathbf{x}_{\mathbf{u}}^{(i)})$ — евклидово расстояние между объектами \mathbf{u} и $\mathbf{x}_{\mathbf{u}}^{(i)}$, вычисленное по матрице объект-признак \mathbf{X} . Будем относить объект \mathbf{u} к тому классу, элементов которого окажется больше среди k его ближайших соседей $\mathbf{x}_{\mathbf{u}}^{(i)}$, $i = 1, \dots, k$:

$$h_{\mathbf{u}}(\mathbf{u}, \mathbf{X}^l, k) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^k [y_{\mathbf{u}}^{(i)} = y].$$

Оптимальное значение параметра k определяется по критерию скользящего контроля с исключением объектов группами: производится T исключений из обучающей выборки произвольных групп \mathbf{X}^{l_t} , $t = 1 \dots T$, размер которых составляет 30 % обучающей выборки. Измеряется величина ошибки классификации на этих группах при фиксированном числе k ближайших соседей. Эта ошибка суммируется для всех таких групп и минимизируется по числу ближайших соседей.

$$S(k, \mathbf{X}^l) = \sum_{t=1}^T \sum_{i=1}^{l_t} [h_{\mathbf{u}_i}(\mathbf{u}_i, \mathbf{X}^l \setminus \{\mathbf{X}^{l_t}\}, k) \neq y_i] \rightarrow \min_{k \in \mathbb{N}} \quad (3)$$

Таким образом, предлагаемая модель метрической классификации f может быть представлена в виде композиции нескольких функций:

$$f(\mathcal{D}, k) = h\left(\rho(\mathcal{D} \times g_{\rho}(Y)), \rho(\mathcal{D}^l \times g_{\rho}(Y)), k\right).$$

При этом, решаемая задача (1) выбора структуры модели классификации имеет вид

$$f(\mathcal{D}, k) = \operatorname{arg min}_{k \in \mathbb{N}} (S(f(\mathcal{D}, k), Y | \mathcal{D}_t)).$$

3 Вычислительный эксперимент

С целью получения оценки ошибки предложенной модели и ее сравнения с моделью классификации из [5] был проведен вычислительный эксперимент. Использовались данные акселерометра мобильного телефона. Данные представляют собой 1260 временных рядов длиной 600 точек, каждый из которых последовательно объединяет три временных ряда: временной ряд ускорения по оси X (200 измерений), оси Y (200 измерений) и оси Z (200 измерений). Выделено шесть типов физической активности: ходьба, бег, сидение, стояние, подъем, спуск. Временные ряды записывались акселерометром, который находился в кармане у человека, выполняющего один из типов физической активности, после чего разделялись на 10-секундные сегменты.

Для предварительной оценки качества классификации описанного алгоритма были построены диаграммы рассеяния для множества признаков X . Пример диаграммы рассеяния для признаков шесть и два приведен на рисунке 2 для всех объектов обучающей выборки. По оси X отложены значения признака два, то есть расстояние до центроида второго класса, а по оси Y — значения шестого признака, то есть расстояние до центроида шестого класса. Из-за плохой визуальной разделимости на диаграммах рассеяния выдвигается предположение о низком качестве классификации объектов, принадлежащих классам “ходьба вниз по лестнице” и “ходьба вверх по лестнице”. Остальные же классы визуально разделимы на этих диаграммах. Предполагается, что на них величина ошибки будет значительно ниже.

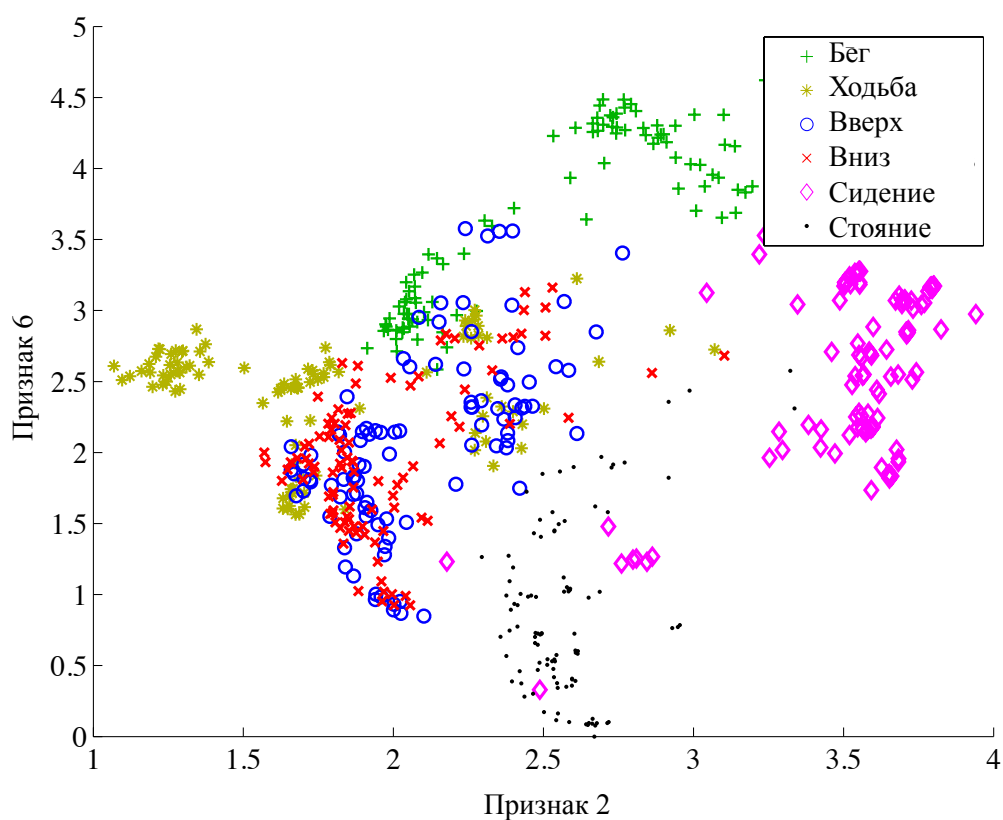


Рис. 2: Диаграмма рассеяния для признаков 6 и 2, по которой выдвигаются предположения о качестве классификации

В таблице 1 показаны средние расстояния между объектами различных классов. Для большинства классов средние расстояния до объектов других классов больше, чем среднее расстояние между объектами самого класса. Выдвигается предположение о низкой величине ошибки на таких классах.

Все вышеизложенные предположения подтверждаются последующими результатами эксперимента.

Таблица 1: Средние расстояния между объектами различных классов

| | Бег | Ходьба | Вверх | Вниз | Сидение | Стояние |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|
| Бег | 1.7221 | 2.7483 | 3.3432 | 3.2631 | 3.4754 | 3.8243 |
| Ходьба | 2.7483 | 1.0665 | 1.4622 | 1.3133 | 2.9301 | 2.1336 |
| Вверх | 3.3432 | 1.4622 | 1.4405 | 1.3760 | 3.1055 | 1.8948 |
| Вниз | 3.2631 | 1.3133 | 1.3760 | 1.2323 | 3.1764 | 1.9166 |
| Сидение | 3.4754 | 2.9301 | 3.1055 | 3.1764 | 1.5114 | 3.2813 |
| Стояние | 3.8243 | 2.1336 | 1.8948 | 1.9166 | 3.2813 | 1.2692 |

При выполнении классификации методом k ближайших соседей при помощи метода скользящего контроля (3) было найдено наилучшее k . На рисунке 3 показано среднее значение функции ошибки S из выражения (1) и ее стандартное отклонение в зависимости от величины k . Имеется минимум при $k = 1$.

В вычислительном эксперименте исследовалась точность модели метрической классификации, описанного в главе 3. Оценивалась точность модели по критерию скользящего контроля (3). В таблице 2 приводится сравнение результатов работы модели, основанной на методе DTW и модели, описанной в [5]. Качество классификации на классе “Ходьба” ухудшилось, в то время, как на классах “вверх” и “вниз” увеличилось. У построенной модели качество классификации выше, чем у модели разделяющей классификации.

Таблица 2: Сравнение эффективности предложенного метрического алгоритма классификации и алгоритма разделяющей классификации на данных [2] (Ход. — Ходьба, Вв. — Вверх, Вн. — Вниз, Сид. — Сидение, Ст. — Стояние)

| | Бег | Ход. | Вв. | Вн. | Сид. | Ст. | Общее |
|---|-----|------|-----|-----|------|-----|-------|
| Метрическая классификация, Точность по критерию скользящего контроля, % | 100 | 83 | 80 | 87 | 99 | 100 | 91 |
| Разделяющая классификация [5], Относительная точность на тестовых данных, % | 98 | 96 | 75 | 72 | 100 | 89 | 88 |

Заключение

В работе была предложена модель метрической классификации. В рамках модели было предложено использование расстояния, вычисленного методом динамического

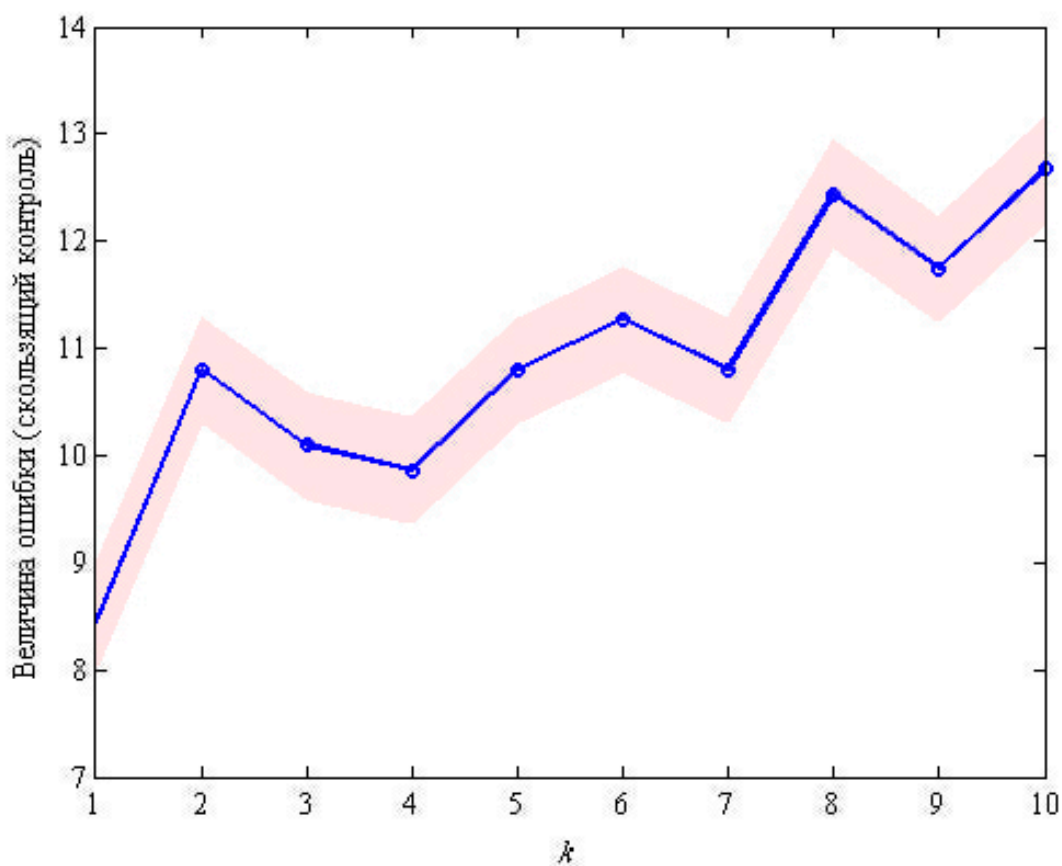


Рис. 3: Зависимость ошибки от величины k

выравнивания временных рядов, использование центроидов для снижения размерности задачи и последующая классификация внутри построенного таким образом признакового пространства. С целью сравнения качества классификации предложенной модели и модели разделяющей классификации был проведен вычислительный эксперимент, демонстрирующий высокое качество классификации модели метрической классификации. Программная реализация алгоритма метрической классификации в среде разработки MatLab находится в свободном доступе [18].

Список литературы

- [1] *Kwapisz J. R., Weiss G. M., Moore S.* Activity recognition using cell phone accelerometers // SIGKDD Explorations, 2010. Vol. 12, No. 2. Pp. 74–82.
- [2] *Kwapisz J. R.* Data from accelerometer // Sourceforge.net: веб-сайт для разработчиков открытого программного обеспечения. 2010.

- http://sourceforge.net/p/mlalgorithms/TSLearning/data/preprocessed_large.csv (дата обращения: 13.03.2015).
- [3] *Hoseini-Tabatabaei S. A., Gluhak A., Tafazolli R.* A survey on smartphone-based systems for opportunistic user context recognition // ACM Computing Surveys (CSUR), 2013. Vol. 45, No. 3, P. 27.
 - [4] *Кузнецов М. П.* Признаковая классификации временных рядов// Sourceforge.net: веб-сайт для разработчиков открытого программного обеспечения. 2014. <https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/TSLearning/doc/TSClassification/TSClassification.pdf> (дата обращения: 10.04.2015).
 - [5] *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015. Т. 9. Вып 1. С. 79–89.
 - [6] *Faloutsos C., Ranganathan M., Manolopoulos Y.* Fast Subsequence Matching in Time-Series Databases // ACM International Conference on Management of Data (SIGMOD), 1994. Pp. 419-429.
 - [7] *Berndt D. J., Clifford J.* Using dynamic time warping to find patterns in time series // Workshop on Knowledge Discovery in Databases (KDD Workshop), 1994. Pp. 359-370.
 - [8] *Keogh E. J., Ratanamahatana C. A.* Exact indexing of dynamic time warping. // Knowl. Inf. Syst., 2005. Vol. 7, No. 3, Pp. 358-386.
 - [9] *Vlachos M., Gunopulos D., Kollios G.* Discovering similar multidimensional trajectories // IEEE International Conference on Data Engineering (ICDE), 2002. Pp. 673-684.
 - [10] *Chen L., Ng R. T.* On the marriage of lp-norms and edit distance // Very Large Data Bases (VLDB), 2004. Pp. 792-803
 - [11] *Chen L., Özsu M. T., Orin V.* Robust and fast similarity search for moving object trajectories // ACM International Conference on Management of Data (SIGMOD), 2005. Pp. 491-502.
 - [12] *Frentzos E., Gratsias K., Theodoridis Y.* Index-based most similar trajectory search // IEEE International Conference on Data Engineering (ICDE), 2007. Pp. 816-825.
 - [13] *Morse M. D., Patel J. M.* An efficient and accurate method for evaluating time series similarity // ACM International Conference on Management of Data (SIGMOD), 2007. Pp. 569-580.
 - [14] *Chen Y., Nascimento M. A., Ooi B. C., Tung A. K. H.* SpADe: On Shape-based Pattern Detection in Streaming Time Series // IEEE International Conference on Data Engineering (ICDE), 2007. Pp. 786-795

- [15] *Scheuermann P., Wang X., Ding H., Trajcevski G., Keogh E.* Querying and mining of time series data: experimental comparison of representations and distance measures // Very Large Data Bases (VLDB), 2008. Pp. 1542-1552.
- [16] *Keogh E. J., Pazzani M. J.* Scaling up Dynamic Time Warping to Massive Datasets // Principles of Data Mining and Knowledge Discovery, 1999. Pp. 1-11.
- [17] *Salvador S., Chan P.* Fastdtw: Toward accurate dynamic time warping in linear time and space // Workshop on Mining Temporal and Sequential Data, 2004. P. 11.
- [18] *Гончаров А. В.* Реализация метрической классификации временных рядов// Sourceforge.net. 2015. <http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group274/Goncharov2015MetricClassification/code/main.m> (дата обращения: 13.05.2015).