ISSN 1684-6400

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

5 (213)
2014

ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

САПР

КОМПЬЮТЕРНАЯ ГРАФИКА

МЕТОДЫ ПРОГРАММИРОВАНИЯ

ОПЕРАЦИОННЫЕ СИСТЕМЫ И СРЕДЫ

ТЕЛЕКОММУНИКАЦИИ И ВЫЧИСЛИТЕЛЬНЫЕ СЕТИ

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

НЕЙРОСЕТИ И НЕЙРОКОМПЬЮТЕРЫ

СТРУКТУРНЫЙ СИНТЕЗ

ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

ОПТИМИЗАЦИЯ И МОДЕЛИРОВАНИЕ

ИТ В ОБРАЗОВАНИИ

ГИС

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

5(213)
2014

ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

Издается с ноября 1995 г.

УЧРЕДИТЕЛЬ
Издательство “Новые технологии”

СОДЕРЖАНИЕ

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Кухаренко Б. Г., Солнцева М. О. Кластеризация управляемых объектов на основе сходства их многомерных траекторий 3
Мохов А. С., Толчеев В. О. Разработка методов высокоточной классификации двуязычных текстовых библиографических документов 8

МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ

Левин В. И. Методология оптимизации в условиях неопределенности методом детерминизации 14

ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ И СЕТИ

Инютин С. А. Особенности вычисления характеристик модулярной величины . .22
Вишнеков А. В., Иванова Е. М., Филиппов В. А. Выбор среды передачи данных при проектировании локальных вычислительных сетей 27

КОДИРОВАНИЕ И ОБРАБОТКА СИГНАЛОВ

Дворников С. В., Цветков В. В., Устинов А. А. Компенсация движения при кодировании подвижных изображений на основе разбиения кодируемого блока кадров на непересекающиеся группы 33

БЕЗОПАСНОСТЬ ИНФОРМАЦИИ

Чичварин Н. В. Выбор методов защиты проектной документации от несанкционированного доступа 41
Долгопятов А. Ю. Восстановление удаленных данных 48
Артюшенко В. М., Аббасова Т. С. Эффективность защиты от внешних помех электропроводных каналов структурированных кабельных систем для передачи высокоскоростных информационных приложений 52

Журнал в журнале
НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ

Аведьян Э. Д., Луганский В. Э. Способы повышения точности заполнения числовых пропусков в таблицах, основанные на модифицированных нейронных сетях СМАС 58
Мышев А. В. Архитектура виртуальной потоковой вычислительной системы на основе информационной модели нейросети. 65

Приложение. Панов А. Д. Технологическая сингулярность, теорема Пенроуза об искусственном интеллекте и квантовая природа сознания

Информация о журнале доступна по сети Internet по адресу **http://novtex.ru/IT.**
Журнал включен в систему Российского индекса научного цитирования.
Журнал входит в Перечень научных журналов, в которых по рекомендации ВАК РФ должны быть опубликованы научные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

© Издательство "Новые технологии", "Информационные технологии", 2014

Главный редактор:
СТЕМПКОВСКИЙ А. Л.,
акад. РАН, д. т. н., проф.

Зам. главного редактора:
ИВАННИКОВ А. Д., д. т. н., проф.
ФИЛИМОНОВ Н. Б., д. т. н., с.н.с.

Редакционный совет:
БЫЧКОВ И. В., акад. РАН, д. т. н.
ЖУРАВЛЕВ Ю. И.,
акад. РАН, д. ф.-м. н., проф.
КУЛЕШОВ А. П.,
акад. РАН, д. т. н., проф.
ПОПКОВ Ю. С.,
чл.-корр. РАН, д. т. н., проф.
РУСАКОВ С. Г.,
чл.-корр. РАН, д. т. н., проф.
СОЙФЕР В. А.,
чл.-корр. РАН, д. т. н., проф.
СОКОЛОВ И. А., акад.
РАН, д. т. н., проф.
СУЕТИН Н. В., д. ф.-м. н., проф.
ЧАПЛЫГИН Ю. А.,
чл.-корр. РАН, д. т. н., проф.
ШАХНОВ В. А.,
чл.-корр. РАН, д. т. н., проф.
ШОКИН Ю. И.,
акад. РАН, д. т. н., проф.
ЮСУПОВ Р. М.,
чл.-корр. РАН, д. т. н., проф.

Редакционная коллегия:
АВДОШИН С. М., к. т. н., доц.
АНТОНОВ Б. И.
БАРСКИЙ А. Б., д. т. н., проф.
ВАСЕНИН В.А., д. ф.-м. н., проф.
ГАЛУШКИН А.И., д. т. н., проф.
ДИМИТРИЕНКО Ю. И., д. ф.-м. н., проф.
ДОМРАЧЕВ В.Г., д. т. н., проф.
ЗАГИДУЛЛИН Р. Ш., к. т. н., доц.
ЗАРУБИН В. С., д. т. н., проф.
ИСАЕНКО Р. О., к. т. н., с.н.с.
КАРПЕНКО А. П., д. ф.-м. н., проф.
КОЛИН К. К., д. т. н., проф.
КУЛАГИН В. П., д. т. н., проф.
КУРЕЙЧИК В. М., д. т. н., проф.
КУХАРЕНКО Б. Г., к. ф.-м. н., доц.
ЛЬВОВИЧ Я. Е., д. т. н., проф.
МИХАЙЛОВ Б. М., д. т. н., проф.
НЕЧАЕВ В. В., к. т. н., проф.
РЯБОВ Г. Г., чл.-корр. РАН, д. т. н., проф.
СОКОЛОВ Б. В., д. т. н., проф.
УСКОВ В. Л., к. т. н. (США)
ФОМИЧЕВ В. А., д. т. н., проф.
ЧЕРМОШЕНЦЕВ С. Ф., д. т. н., проф.
ШИЛОВ В. В., к. т. н., доц.

Редакция:
БЕЗМЕНОВА М. Ю.
ГРИГОРИН-РЯБОВА Е. В.
ЛЫСЕНКО А. В.
ЧУГУНОВА А. В.

УДК 519.233

Б. Г. Кухаренко, канд. физ.-мат. наук, ст. науч. сотр., вед. науч. сотр.,
Институт машиноведения РАН, г. Москва, e-mail: kukharenko@imash.ru,
М. О. Солнцева, аспирант, Московский физико-технический институт (ГУ),
e-mail: solnceva.chalei@gmail.com

Кластеризация управляемых объектов на основе сходства их многомерных траекторий

Для кластеризации многомерных траекторий применяется метод полиномиальных регрессий с обучением параметров посредством алгоритма ожидания и максимизации правдоподобия. Особенностью метода является одновременное выравнивание траекторий в многомерном пространстве и во времени. Эффективность метода кластеризации демонстрируется на примере анализа траекторий движения самолетов в воздушном пространстве аэропорта.

Ключевые слова: *анализ данных, многомерные траектории, кластеризация, полиномиальная регрессия, алгоритм ожидания и максимизации правдоподобия*

B. G. Kukharenko, M. O. Solntseva

Clustering Objects under Control by Similarity of Their Multidimensional Trajectories

For clustering multidimensional trajectories polynomial regression method with parameter leaning by the Expectation-Maximization algorithm is in use. The method based on polynomial regression is characterized by the joint clustering and continuous alignment of curve sets in time and space. Efficiency of the clustering method is demonstrated by analysis of flight tracks in airport space.

Keywords: *data analysis, multidimensional trajectories, clustering, polynomial regression, Expectation-Maximization algorithm*

Введение

Современное развитие приборов мобильной связи и систем теле- и видеонаблюдения, в том числе использование космических спутников, способствует накоплению громадного количества данных о траекториях перемещения людей, наземного транспорта, морских, воздушных судов и т. п. Классификация траекторий движущихся объектов, т. е. построение моделей для предсказания классов таких объектов (согласно их траекториям и различным дополнительным характеристикам), используется в ряде практических приложений для анализа транспортных сетей [1, 2], создания военных систем наблюдения [3], морского пограничного контроля [4], оптимизации загруженности взлетно-посадочной полосы и обеспечения безопасности воздушного пространства в крупных аэропортах [5]. Задача кластеризации траекторий движения возникает и при управлении мультиробототехническими системами. Она обусловлена необходимостью предсказания движения для организации взаимодействия объектов системы на основе выделяемых паттернов в траекториях движения [6]. Кластеризация объектов робототехнических систем также упрощает контроль их согласованного перемещения [7, 8].

Чтобы исключить в любой рассматриваемой системе нежелательную конкуренцию (столкновения) между объектами, совершающими движение по сходным траекториям, следует разнести их во времени и (или) в пространстве. В этом случае также выполняется предварительная кластеризация движущихся объектов системы в соответствии с их траекториями. Для кластеризации кривых, представляющих траектории движения, применяют различные методы, из которых наиболее эффективен метод локальной полиномиальной регрессии [9]. Для

кластеризации траекторий при обучении полиномиальной регрессии в дальнейшем используется итеративный алгоритм ожидания и максимизации правдоподобия (EM-алгоритм) [10]. Этот объединенный подход обеспечивает выравнивание кривых, осуществляемое в координатном пространстве и во времени одновременно. Задача выравнивания кривых с использованием модели априорных вероятностей, заданных на множестве возможных выравниваний, решается EM-алгоритмом, что формализует так называемый подход Прокруста для анализа кривых [11]. Априорные вероятности выравнивания интегрируются в модель смеси конечного числа распределений, на основе которой выполняется кластеризация. В настоящей работе для выполнения одновременной кластеризации и выравнивания кривых используется полиномиальная модель.

Ранее модель смеси гауссовых распределений с обучением посредством EM-алгоритма использовалась для совместной кластеризации и выравнивания при корректировке изображений, которые подвергались различным линейным преобразованиям [12]. Однако такая модель рассматривает преобразования в дискретном пространстве пикселей, в то время как подход, применяемый в настоящей работе, направлен на моделирование кривых и допускает произвольное и непрерывное их выравнивание в пространстве и во времени. Эффективность такого подхода демонстрируется на примере анализа данных радаров международного аэропорта г. Сан-Франциско, находящихся в свободном доступе на сайте https: //c3.nasa.gov/dashlink/resources/132/

Модель одновременной кластеризации и выравнивания одномерных временных рядов

Рассмотрим одновременные кластеризацию и выравнивание одномерных временных рядов, т. е. векторов переменной длины, представляющих координатные временные зависимости. Каждый вектор $\mathbf{z}_i \in \mathbb{R}^{N_i \times 1}$, $i = 1, 2, \ldots$, состоит из последовательности измерений координатной зависимости $z_i = z_i(t)$ в моменты времени $\mathbf{t}_i \in \mathbb{R}^{N_i \times 1}$. В работе [13] модели смеси регрессий эффективно используются для кластеризации одномерных векторов переменной длины. Вектор $\mathbf{z}_i$ моделируется регрессионной моделью

$$\mathbf{z}_i = \mathbf{T}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{1}$$

где $\boldsymbol{\beta}$ — вектор коэффициентов регрессии размерности $(q + 1) \times 1$; $\boldsymbol{\varepsilon}_i$ — гауссов шум с нулевым средним; $\mathbf{T}_i$ — регрессионная матрица. Матрица $\mathbf{T}_i$ зависит от типа используемой регрессионной модели. В случае полиномиальной регрессии $\mathbf{T}_i$ имеет вид стандартной матрицы Вандермонта

$$\mathbf{T}_i = \begin{bmatrix} 1 & t_i[1] & (t_i[1])^2 & \ldots & (t_i[1])^q \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & t_i[N_i] & (t_i[N_i])^2 & \ldots & (t_i[N_i])^q \end{bmatrix}. \tag{2}$$

В основе модели одновременной кластеризации и выравнивания лежит модель смеси регрессий, в которую вводятся четыре независимых параметра преобразований выравнивания и масштабирования во времени и пространстве $\{\Phi_i\} = \{a_i, b_i, c_i, d_i\}$ (параметры a_i и b_i описывают масштабирование и сдвиг во времени, а параметры c_i и d_i — масштабирование и смещение в пространстве измерений). Полиномиальная регрессия для одномерного случая имеет вид

$$\mathbf{z}_i = c_i\boldsymbol{\Upsilon}_i\boldsymbol{\beta}_k + d_i + \boldsymbol{\varepsilon}_i, \tag{3}$$

где матрица $\boldsymbol{\Upsilon}_i$ получается из $\mathbf{T}_i$ (2) подстановкой $\mathbf{t}_i \to a_i\mathbf{t}_i - b_i$; $\boldsymbol{\beta}_k$ определяет модель регрессии для k-го кластера ($k = \overline{1, K}$); $\boldsymbol{\varepsilon}_i$ — гауссов шум с нулевым средним и дисперсией $\sigma_k^2\mathbf{I}$. Поэтому распределение плотности условной вероятности имеет вид

$$p_k(\mathbf{z}_i|a_i, b_i, c_i, d_i) = \mathcal{N}(\mathbf{z}_i|c_i\boldsymbol{\Upsilon}_i\boldsymbol{\beta}_k + d_i, \boldsymbol{\sigma}_k^2\mathbf{I}). \tag{4}$$

Плотность вероятности для кривой $\mathbf{z}_i$ однозначно задается соответствующим множеством параметров $\{\Phi_i\}$, которые подлежат определению. Задача кластеризации кривых решается как стандартная задача оценки значений скрытых переменных. Каждый из параметров преобразования в формулах (3) и (4) рассматривается как характерная для $\mathbf{z}_i$ случайная переменная с заранее известным распределением вероятности для кластера. Параметры преобразования и параметры модели оцениваются одновременно посредством EM-алгоритма.

Априорные распределения вероятностей для параметров преобразования

Априорные распределения вероятностей для параметров преобразования выбираются таким образом, чтобы тождественное преобразование являлось наиболее вероятным. С учетом этого эффективной априорной вероятностью является Гауссово распределение $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ со средним $\boldsymbol{\mu}$ и дисперсией $\boldsymbol{\sigma}^2$. Поэтому априорные распределения плотности вероятности для параметров преобразования времени определяются как

$$a_i \sim \mathcal{N}(1, r_k^2),\ b_i \sim \mathcal{N}(0, s_k^2), \tag{5}$$

и априорные распределения вероятности для параметров преобразования координаты задаются как

$$c_i \sim \mathcal{N}(1,\ u_k^2),\ d_i \sim \mathcal{N}(1,\ v_k^2). \qquad (6)$$

В формуле (6) параметры дисперсии u_k^2 и v_k^2 зависят от кластера. Однако любое подмножество этих параметров может быть выравнено ("сшито") между кластерами, если это требуется для конкретного приложения. Отметим, что априорные распределения вероятности технически допускают отрицательное масштабирование во времени и в пространстве измерений. Хотя этот результат не типичен, можно задать другие априорные распределения вероятности, например, логарифмически нормальные, чтобы не допустить отрицательного масштабирования. Следует заметить, что дисперсии для параметров априорных вероятностей выводятся на основе данных, полученных в результате работы EM-алгоритма. Ниже модель (3)—(6) совместной кластеризации и выравнивания кривых обобщается на случай многомерного пространства измерений.

Кластеризация и выравнивание многомерных кривых

Рассмотрим выравнивание кривых в многомерном пространстве. Ранее предполагалось, что вектор $\mathbf{z}_i \in \mathbb{R}^{N_i \times 1}$ состоит из последовательности измерений одномерной координатной зависимости в моменты времени $\mathbf{t}_i \in \mathbb{R}^{N_i \times 1}$. Однако во многих приложениях такие зависимости от времени являются многомерными. Таким образом, каждому моменту времени $t_i[j], j = \overline{1, N_i}$, соответствует многомерный вектор размерности D. Обозначим многомерные кривые как $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$, полученные в результате измерений в моменты времени $\mathbf{t}_i \in \mathbb{R}^{N_i \times 1}$. Тогда матрица $\mathbf{Z}_i = \{Z_i[j\ ;\ l], j = \overline{1, N_i},\ l = \overline{1, D}\ \}$ состоит из D столбцов, таких что каждый l-й столбец $\mathbf{z}_i^{(l)} = \{Z_i[j;\ l],\ j = \overline{1, N_i}\},\ l = \overline{1, D}$ содержит последовательность измерений l-й одномерной координатной зависимости для i-й рассматриваемой переменной. То есть столбец $\mathbf{z}_i^{(l)}$, соответствующий одномерному вектору $\mathbf{z}_i \in \mathbb{R}^{N_i \times 1}$ в формуле (1), оказывается вложенным в матрицу многомерной кривой $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$.

Регрессионные модели выравнивания в многомерном пространстве. Для многомерной кривой $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$ вектор измерения каждой координаты описывается независимой регрессионной моделью

$$\mathbf{z}_i^{(l)} = \mathbf{T}_i \boldsymbol{\beta}_{kl} + d_{il} + \boldsymbol{\varepsilon}_i;\ d_{il} \propto \mathcal{N}(0,\ \nu_{kl}^2);$$
$$\boldsymbol{\varepsilon}_i \propto \mathcal{N}(\mathbf{0}_{N_i \times 1},\ \sigma_{kl}^2 \mathbf{I}_{N_i \times N_i}), \qquad (7)$$

где $\mathbf{T}_i$ — матрица Вандермонта (2). Матрица $\boldsymbol{\beta}_{kl}$ задает коэффициенты регрессии для l-го измерения (т. е. коэффициенты регрессии для l-го столбца $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$); d_{il} задает смещение для l-го измерения; $\mathbf{0}_{N_i \times 1}$ — вектор с нулевыми компонентами размерности $N_i \times 1$ и $\mathbf{I}_{N_i \times N_i}$ — тождественная матрица размерности $N_i \times N_i$. Использование параметров ν_{kl}^2, σ_{kl}^2 позволяет рассматривать дисперсию по каждому измерению независимо.

На основе модели (7) плотность вероятности многомерной кривой $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$ и число параметров D смещения $\{d_{il},\ l = \overline{1, D}\ \}$ определяется следующим образом:

$$p(\mathbf{Z}_i,\ d_{i1},\ ...,\ d_{iD}) =$$
$$= \prod_{l=1}^{D} \mathcal{N}(\mathbf{z}_i^{(l)} | \mathbf{T}_i \boldsymbol{\beta}_{kl} + d_{il},\ \sigma_{kl}^2 \mathbf{I}_{N_i \times N_i}) \mathcal{N}(d_{il} | 0,\ \nu_{kl}^2). \qquad (8)$$

Плотность вероятности (8) учитывает два необходимых условия: во-первых, все координаты пространства кривых $\mathbf{Z}_i \in \mathbb{R}^{N_i \times D}$ и, во-вторых, предполагается, что для каждого измерения l существует собственное множество параметров смещения $\{\Phi_i\}_l$. Далее предполагается, что эти два условия всегда выполняются.

Плотность безусловного распределения (компонент многомерной случайной величины) $p(\mathbf{Z}_i)$ представляется в виде произведения $p(\mathbf{Z}_i) = \prod_{l=1}^{D} p(\mathbf{z}_i^{(l)})$, поэтому логарифм правдоподобия $\mathbb{Z} = \{\mathbf{Z}_i,\ i = \overline{1, N}\}$ имеет вид

$$\log(p(\mathbb{Z})) = \sum_{i=1}^{N} \log(p(\mathbf{Z}_i)) =$$
$$\sum_{i=1}^{N} \sum_{l=1}^{D} \log(\int p(\mathbf{z}_i^{(l)} | d_{il}) p(d_{il}) dd_{il}). \qquad (9)$$

Интегрирование в формуле (9) выполняется аналитически, что приводит к следующему виду логарифма правдоподобия:

$$\log(p(\mathbb{Z})) = \\ = \sum_{i=1}^{N}\sum_{l=1}^{D}\log(\mathcal{N}(\mathbf{z}_i^{(l)}|\mathbf{T}_i\boldsymbol{\beta}_{kl}, \mathbf{1}_{N_i\times N_i}\nu_{kl}^2 + \sigma_{kl}^2\mathbf{I}_{N_i\times N_i})), \quad (10)$$

где $\mathbf{1}_{N_i\times N_i}$ — единичная матрица размерности $N_i \times N_i$.

Регрессионные модели выравнивания многомерных кривых во времени. Поскольку независимые координаты многомерного пространства смещаются и масштабируются независимо, для выравнивания в пространстве рассматриваются D отдельных параметров преобразования. Однако изменение во времени каждой координаты траектории происходит в одном и том же временном масштабе, следовательно, параметры преобразования времени должны быть распределены одинаково по всем D измерениям. Поэтому каждому из векторов $\mathbf{z}_i^{(l)} \in \mathbb{R}^{N_i\times 1}$ соответствует единственный параметр b_i. Тогда условная плотность вероятности $\mathbf{Z}_i \in \mathbb{R}^{N_i\times D}$ определяется как

$$p(\mathbf{Z}_i|b_i) = \prod_{l=1}^{D} p(\mathbf{z}_i^{(l)}|b_i) = \\ = \prod_{l=1}^{D}\mathcal{N}(\mathbf{z}_i^{(l)}|\boldsymbol{\Upsilon}_i\boldsymbol{\beta}_{kl}, \sigma_{kl}^2\mathbf{I}_{N_i\times N_i}), \quad (11)$$

где используется одно b_i для всех $l = \overline{1, D}$ и матрица $\boldsymbol{\Upsilon}_i$ получается из $\mathbf{T}_i$ (2) подстановкой $\mathbf{t}_i \to \mathbf{t}_i - b_i$. Условная плотность вероятности разлагается на множители, а безусловная плотность вероятности $p(\mathbf{Z}_i)$ не разлагается, поскольку различные измерения в пространстве оказываются связанными через параметр смещения времени b_i. Следовательно, для логарифма правдоподобия $\mathbb{Z} = \{\mathbf{Z}_i, i = \overline{1, N}\}$ имеем

$$\log(p(\mathbb{Z})) = \\ = \sum_{i=1}^{N}\log(p(\mathbf{Z}_i)) = \sum_{i=1}^{N}\log(\int p(b_i)\prod_{l=1}^{D} p(\mathbf{z}_i^{(l)}|b_i)db_i). \quad (12)$$

Под знаком интеграла в формуле (12) находится произведение вероятностей по всем измерениям пространства, что приводит к сложным вычислениям. Однако с помощью методов Монте-Карло можно вычислить аппроксимацию логарифма правдоподобия (12) следующим образом:

$$\log(p(\mathbb{Z})) \approx \sum_{i=1}^{N}\log\left(\sum_{m=1}^{M}\prod_{l=1}^{D} p(\mathbf{z}_i^{(l)}|b_i^{(m)})\right) - N\log(M), \quad (13)$$

где

$$b_i^{(m)} \propto \mathcal{N}(0, \eta^2),\ m = \overline{1, M}. \quad (14)$$

EM-алгоритм

Сложность EM-алгоритма, обеспечивающего одновременное обучение параметров модели и преобразования, является линейной функцией от полного числа точек $\sum_{i=1}^{N} N_i$ многомерных траекторий $\mathbf{Z}_i$ [10]. Пусть $\boldsymbol{\pi}_i$ — принадлежность $\mathbf{Z}_i$ к некоторому кластеру. Параметры $\{\Phi_i\}$ и принадлежности к кластеру рассматриваются как скрытые переменные. В таком случае логарифм правдоподобия для полного набора данных определяется как логарифм совместного правдоподобия множества $\mathbb{Z} = \{\mathbf{Z}_i, i = \overline{1, N}\}$ и скрытых переменных $\Phi = \{\Phi_i, \pi_i\}$, что в соответствии с формулой (11) может быть записано в виде суммы (по всем N кривым) логарифма от произведения веса кластера α_{π_i} и совместного распределения вероятности (8), зависящего от кластера

$$L_c = \sum_{i=1}^{N}\sum_{l=1}^{D}\log(\alpha_{\pi_i} p_{\pi_i}(\mathbf{z}_i^{(l)}|\Phi_i)p_{\pi_i}(\Phi_i)). \quad (15)$$

На E-шаге оценивается распределение вероятности $p(\Phi_i, \pi_i|\mathbf{Z}_i)$ и затем используется в качестве следующего ожидаемого распределения в (15). На следующей итерации это ожидаемое распределение используется на M-шаге для оценки параметров модели в $p_{\pi_i}(\mathbf{z}_i^{(l)}|\Phi_i)$.

Численный эксперимент

Описываемый в настоящей работе подход к кластеризации кривых тестируется на реальных данных радара TRACON (Terminal Radar Approach Control), регистрирующего траектории полетов воздушных судов над заливом Сан-Франциско (данные находятся в открытом доступе на сайте https://c3.nasa.gov/dashlink/resources/132/). Рассматривается область воздушного пространства над тремя крупными аэропортами: Окленда, Сан-Франциско и Сан-Хосе (Oakland, San Francisco and San Jose International Airports). Это пространство представляет собой цилиндр радиуса 80 км с центром над аэропортом Окленда и высотой 6 км. Данные содержат трехмерные координаты и абсолютные скорости воздушных судов через равные интервалы времени (5 с) с момента обнаружения радаром и до самой нижней регистрируемой радарами высоты. Помимо основной информации в записях указываются дополнительные

сведения о типе совершаемой операции (прибытии или отправлении), о пункте вылета и назначения и др.

В настоящей работе анализируются траектории первых 30 самолетов, которые приземлились в аэропорту 1 января 2006 г. Чтобы их случайные маневры до начала снижения не искажали общей тенденции движения, анализируются только 160 последних точек каждой траектории. Разница между моментами времени последовательной регистрации самолета (~5 с) определяется угловой скоростью вращения радара. Начало координат (0, 0, 0) совпадает с положением радаров.

На рис. 1 (см. третью сторону обложки) показано трехмерное представление для 30 анализируемых траекторий самолетов (идущих на посадку в трех международных аэропортах, в пространстве над заливом Сан-Франциско). В численном эксперименте эти траектории разделяются на пять кластеров. В каждом из пяти кластеров, с помощью описанной выше регрессионной модели (11) выравнивания кривых в многомерном пространстве, определяется собственный тренд.

На рис. 2 (см. третью сторону обложки) показано изменение координат самолетов отдельно по осям (x, y, z) в соответствии с последовательностью моментов времени регистрации радарами. Как и на рис. 1, число временных точек регистрации и интервалы между ними (5 с) — одинаковые для всех траекторий движения самолетов. Разница в реальном времени для самолетов, начинающих снижение, не учитывается, поэтому по горизонтальной оси на рис. 2 отложен индекс моментов времени регистрации радарами. Это позволяет увидеть сходство формы траекторий самолетов, принадлежащих одному кластеру при изменении каждой координаты во времени. Линия тренда (жирная линия на рис. 2) представляет собой некоторую обобщенную форму траекторий кривых в кластере. Например, на рис. 2, *а* два кластера имеют почти линейную форму линии тренда. На рис. 2, *б* для этих же самых кластеров форма линии тренда напоминает прямой угол. На рис. 2, *в* можно заметить линии тренда с выраженной s-образной формой кривой. Несмотря на то, что снижение самолетов происходит не синхронно в реальном времени, их сходные траектории отражают типичные маршруты посадки, когда самолеты выстраиваются в "караван", ожидая своей очереди приземления на посадочную полосу. Рис. 2 является иллюстрацией эффективности метода кластеризации кривых, представленного в настоящей работе.

На рис. 3 (см. третью сторону обложки) данные по кластеризации траекторий самолетов, совершающих посадку (см. рис. 1), показаны в трехмерном пространстве. Как видно на рис. 3, каждый кластер представляется своим "посадочным" паттерном, который проявляет наибольшее сходство со всеми кривыми, вошедшими в кластер, что достигается благодаря поиску максимального значения логарифма правдоподобия (13).

Заключение

В настоящей работе метод одновременной кластеризации и выравнивания кривых во времени и пространстве демонстрируется на примере кластеризации траекторий самолетов, идущих на посадку в зоне крупного международного аэропорта. Эффективность кластеризации траекторий в трехмерном пространстве достигается благодаря использованию модели полиномиальной регрессии с обучением параметров посредством ЕМ-алгоритма.

Список литературы

1. **Кухаренко Б. Г., Солнцева М. О.** Принцип минимальной длинны при анализе графов с разреженными матрицами смежности в задачах кластеризации их узлов // Информационные технологии. 2013. № 7. С. 37—42.
2. **Солнцева М. О., Кухаренко Б. Г.** Применение методов кластеризации узлов на графах с разреженными матрицами смежности в задачах логистики // Труды МФТИ. 2013. Т. 5, № 3 (19). С. 75—83.
3. **Zheng Y., Zhou X.,** eds. Computing with Spatial Trajectories. New York; Dordrecht, Heidelberg, London: Springer. 2011.
4. **Greidanus H., Kourti N.** Findings of the DECLIMS project — Detection and classification of marine traffic from space / Sawaya-Lacoste H., Ouwehand L., eds. // Proceedings of SEASAR 2006 "Advances in SAR Oceanography from Envisat and ERS Missions". Frascati, Italy. 23—26 January 2006. Noordwijk: The European Space Agency. 2006. P. 25—33.
5. **Gariel M., Srivastava A. N., Feron E.** Trajectory clustering and an application to airspace monitoring // IEEE Transactions on Intelligent Transportation Systems. 2011. V. 12, Is. 4. P. 1511—1524.
6. **Sung C., Feldman D., Rus D.** Trajectory clustering for motion prediction, Intelligent Robots and Systems (IROS) // Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012). Vilamoura-Algarve, Portugal: IEEE. 2012. P. 1547—1552.
7. **Солнцева М. О., Кухаренко Б. Г.** Сокращение неинформативной части панорамного кадра по методу швов при управлении группами объектов // Сб. науч. тр. МФТИ "Математические и информационные модели управления". М.: МФТИ. 2013. С. 91—97.
8. **Кухаренко Б. Г., Солнцева М. О.** Использование методов сокращения фона при сегментировании телеметрических изображений для идентификации групп объектов // Информационные технологии. 2014. № 2. С. 3—8.
9. **Fan J., Gijbels I.** Local Polynomial Modelling and its Applications. London: Chapman and Hall. 1996.
10. **Gaffney S., Smyth P.** Joint probabilistic curve clustering and alignment // Saul L., Weiss Y., Bottou L., eds. Proceedings of Neural Information Processing Systems (NIPS 2004). December 13—18, 2004, Vancouver, British Columbia, Canada. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press. 2005. V. 17. P. 473—480.
11. **Ramsay J. O., Silverman B. W.** Functional Data Analysis. New York: Springer-Verlag. 1997.
12. **Frey B. J., Jojic N.** Transformation-invariant clustering using the EM algorithm // IEEE Transactions on PAMI. January 2003. V. 25, N 1. P. 1—17.
13. **Gaffney S., Smyth P.** Trajectory clustering with mixtures of regression models / Chaudhuri S., Madigan D., eds. // Proceedings of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 15—18 1999. New York: ACM Press. 1999. P. 63—72.

Рисунки к статье
Б.Г. Кухаренко,
М.О. Солнцевой
«Кластеризация управляемых объектов на основе сходства их многомерных траекторий»

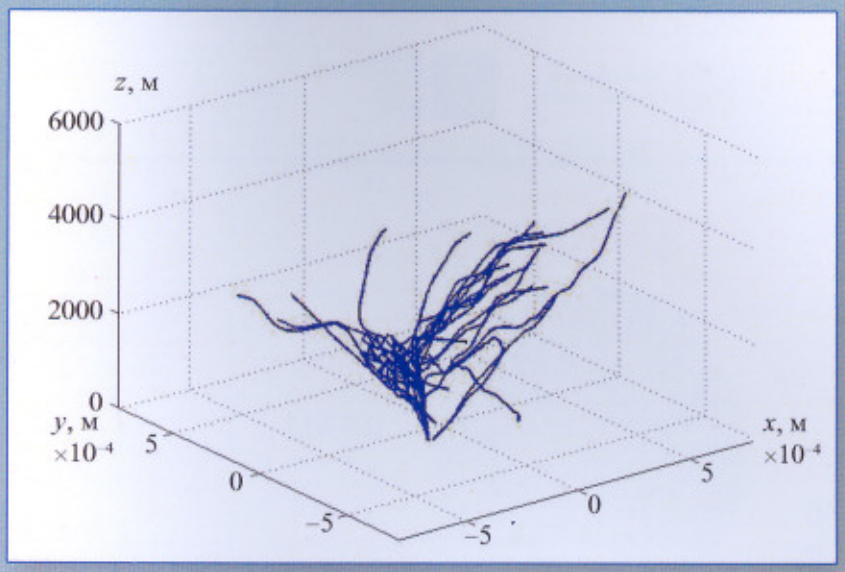

Рис. 1. Изображение для 30 траекторий самолетов, идущих на посадку

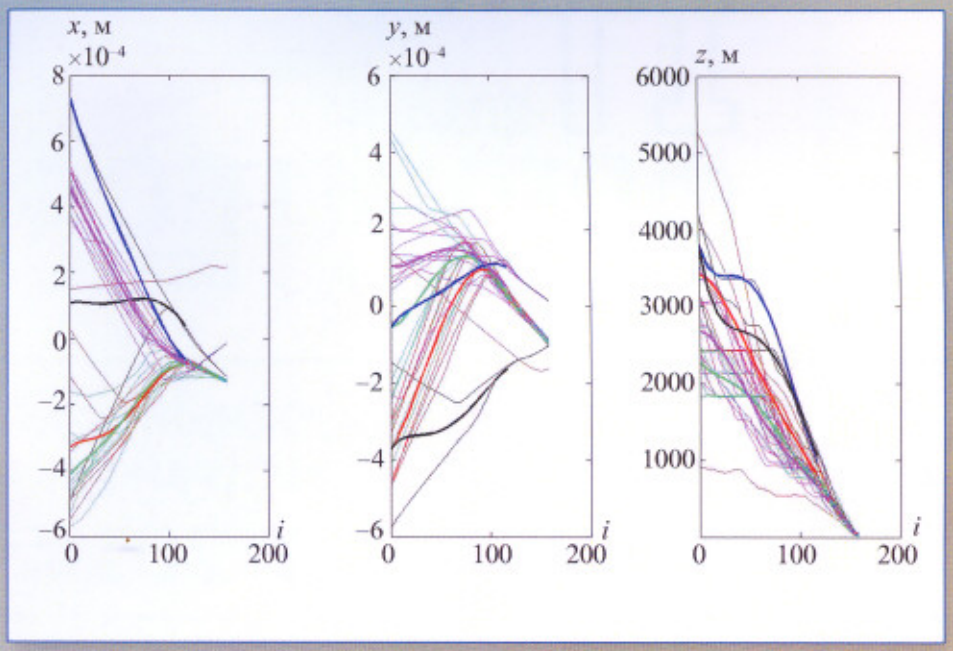

Рис. 2. Проекции траекторий самолётов и линий трендов (жирные линии) для пяти кластеров на координаты трехмерного пространства в зависимости от индекса момента времени регистрации радарами

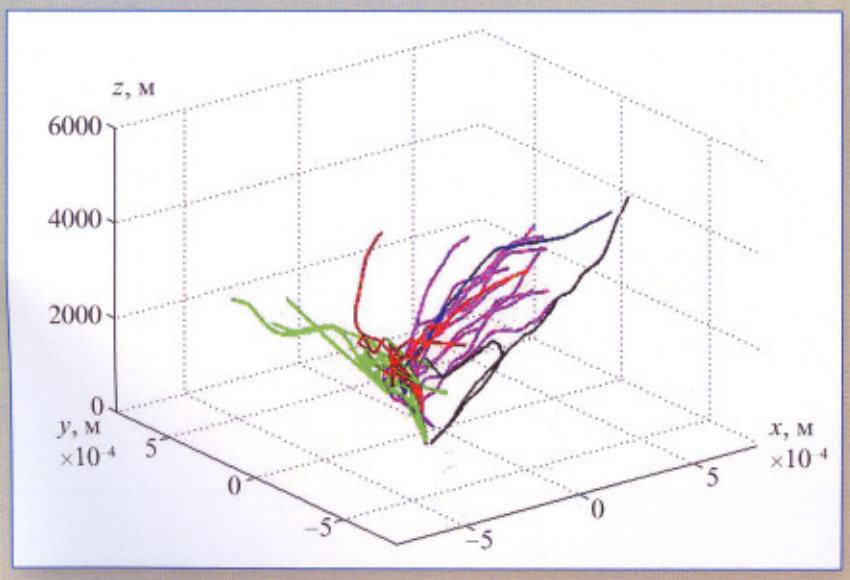

Рис. 3. Результат кластеризации траекторий самолётов в трехмерном пространстве