УДК 519.246

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-Чалей, аспирант,
Московский физико-технический институт (ГУ), e-mail: solnceva.chalei@gmail.com

Спектральный метод с использованием полярной кривизны для анализа результатов кластеризации многомерных траекторий

Геометрический метод используется для получения модели кластеризации в виде объединения аффинных подпространств. Подобие набора векторов оценивается многомерным тензором близости векторов, который развертывается в матрицу близости (подобия) векторов и анализируется спектральным методом. В качестве примера рассматривается кластеризация траекторий движения самолетов при посадке в аэропорту. Оценка близости в кластере по мере косинуса обнаруживает потенциально посторонние траектории полета самолетов в зоне риска.

Ключевые слова: *анализ данных, многомерные траектории, модель объединения аффинных подпространств, полярная кривизна, спектральная кластеризация*

Введение

При анализе наборов существенно многомерных векторов нельзя использовать методы визуализации, обеспечивающие оценку числа кластеров. Поэтому все алгоритмы кластеризации имеют ту же проблему, что и алгоритм K-средних. Примером является кластеризация в пространстве многомерных векторов $\mathbb{R}^D$ (где $D \gg 1$) на основе генеративных моделей, в которых алгоритм ожидания-максимизации правдоподобия — Expectation-Maximization (EM) используется для обучения смеси распределений. Во-первых, чтобы оценить это параметрическое распределение, необходимо сделать предположение о Гауссовом распределении в каждом определяемом кластере. Во-вторых, логарифм правдоподобия может иметь много локальных минимумов, и, следовательно, требуются многочисленные запуски, чтобы получить приемлемое решение. Позже алгоритм K-средних был обобщен, чтобы определять K наилучшим образом аппроксимирующих d-мерных аффинных подпространств для набора векторов в $\mathbb{R}^D$, т. е. прототипом для набора векторов становится аппроксимирующее аффинное подпространство, а не центроид, как в алгоритме K-средних.

Альтернативой является применение спектральных методов кластеризации, в которых используются главные собственные вектора матрицы близости (подобия), основанной на Евклидовом расстоянии между многомерными векторами. Спектральные методы успешно применяют к задачам сегментирования изображений (размерность пространства характеристик для пикселей изображения $\leqslant 7$), но их применимость ограничена парными мерами близости (подобия) при формировании матрицы близости (подобия). Однако геометрические задачи требуют анализа выборки более двух многомерных векторов, чтобы оценить их меру подобия. При решении таких задач определяется вероятность принадлежности к одному и тому же кластеру для набора векторов (а не пары), что приводит к многомерному тензору близости (подобия). Ниже показано, что спектральный метод кластеризации с использованием многомерного тензора близости (подобия) векторов обеспечивает точную кластеризацию для существенно многомерных данных, представляющих траектории движения самолетов при посадке в аэропорту (данные в открытом доступе на сайте https://c3.nasa.gov/dashlink/resources/132/).

1. Спектральная кластеризация с тензором близости на основе полярной кривизны

Задача гибридного линейного моделирования (*hybrid linear modeling*) предполагает, что набор данных (многомерных векторов) достаточно хорошо аппроксимируется объединением аффинных подпространств (*flats*), и необходимо одновременно оценить параметры каждого из аффинных пространств и ассоциацию этих многомерных векторов с аффинными подпространствами [1]. Аффинное d-мерное

подпространство является подмножеством пространства векторов $\mathbb{R}^D$ и характеризуется решением линейной системы уравнений $\mathbb{F} = \{\mathbf{z} | \mathbf{z} \in \mathbb{R}^D, \mathbf{F}^{\mathrm{T}}\mathbf{z} = \boldsymbol{\gamma}\}$, где $\mathbf{F} \in \mathbb{R}^{D\times(D-d)}$, $\boldsymbol{\gamma} \in \mathbb{R}^{1\times(D-d)}$ (например, 0-мерное аффинное подпространство (0-flat) — точка; 1-мерное (1-flat) — плоскость; (D-1)-мерное ((D-1)-flat) — гиперплоскость). В настоящей работе рассматривается специальный случай гибридного линейного моделирования, когда все аффинные подпространства имеют одинаковую размерность $d \geqslant 0$ (*d-flats clustering*) [2]. Используется определенный в работе [3] многомерный тензор аффинности (близости) (*affinity tensor*) для набора векторов и алгоритм спектральной кластеризации [4, 5]. Для каждых (d + 2) векторов из набора данных назначается аффинная мера и в результате формируется (многомерный) тензор аффинности (близости) порядка (d + 2). Развертывание этого тензора близости в матрицу близости (подобия) обеспечивает применение спектральной кластеризации [4, 5].

Пусть d и D — целые числа, такие, что $0 \leqslant d < D$. Для каждых (d + 2) различных векторов-столбцов $\mathbf{z}_1, ..., \mathbf{z}_{d+2} \in \mathbb{R}^D$, $V_{d+1}(\mathbf{z}_1, ..., \mathbf{z}_{d+2})$ обозначает объем выпуклой оболочки, образованной (d + 1) векторами — симплекса (d + 1)-го порядка ((d + 1)-мерного обобщения треугольника или (d + 1)-мерного тетраэдра) [6]. В каждой вершине $\mathbf{z}_i$ полярный синус определяется как

$$\mathrm{psin}_{\mathbf{z}_i}(\mathbf{z}_1, ..., \mathbf{z}_{d+2}) = \frac{(d+1)!\, V_{d+1}(\mathbf{z}_1, ..., \mathbf{z}_{d+2})}{\prod\limits_{j=1,\, d+2,\, j\neq i} \|\mathbf{z}_j - \mathbf{z}_i\|},$$
$$i = \overline{1, d+2}. \qquad (1)$$

Пусть diam($\mathbb{Z}$) обозначает диаметр набора векторов $\mathbb{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_{d+2}\}$. Полярная кривизна (*polar curvature*) набора из (d + 2) векторов определяется в работах [2, 7] как

$$c_{\mathrm{p}}(\mathbf{z}_1, ..., \mathbf{z}_{d+2}) =$$
$$= \mathrm{diam}(\{\mathbf{z}_1, ..., \mathbf{z}_{d+2}\}) \left(\sum_{i=1}^{d+2} (\mathrm{psin}_{\mathbf{z}_i}(\mathbf{z}_1, ..., \mathbf{z}_{d+2}))^2 \right)^{1/2}. \qquad (2)$$

При d = 0 полярная кривизна совпадает с Евклидовым расстоянием. Введем обозначения для набора индексов $I = \{1, ..., 2d\}$ и матрицы из векторов-столбцов $\mathbf{Z}_I = [\mathbf{z}_1 ... \mathbf{z}_{2d}]$. С учетом (1) c_{p} (2) принимает вид

$$c_{\mathrm{p}}(\mathbf{z}_1, ..., \mathbf{z}_{d+2}) =$$
$$= \max_{i,\, j \in I} \|\mathbf{z}_j - \mathbf{z}_i\| \left(\frac{1}{d+2} \sum_{j \in I} \frac{\det(\mathbf{Z}_I^{\mathrm{T}}\mathbf{Z}_I + \mathbf{1})}{\prod\limits_{k \in I,\, k \neq j} \|\mathbf{z}_j - \mathbf{z}_i\|^2} \right)^{1/2}.$$

Числитель $\det(\mathbf{Z}_I^{\mathrm{T}}\mathbf{Z}_I + \mathbf{1})$ является, с точностью до множителя, квадратом объема симплекса (d + 1)-го порядка, сформированного (d + 2) векторами $\{\mathbf{z}_1, ..., \mathbf{z}_{d+2}\}$. Следовательно, полярную кривизну можно рассматривать, как объем симплекса (d + 1)-го порядка, нормированного в каждой вершине, усредненного по вершинам и затем масштабированного диаметром симплекса (d + 1)-го порядка. Если (d + 2) векторов выбраны из одного и того же аффинного подпространства, то полярная кривизна $c_{\mathrm{p}} \approx 0$ и, следовательно, аффинность (близость) ≈ 1. Вместе с тем, когда вектора выбраны из объединения аффинных подпространств, то полярная кривизна велика и аффинность (близость) ≈ 0.

Алгоритм K аффинных подпространств разделяет набор данных $\mathbb{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\} \subset \mathbb{R}^D$ на K поднаборов (кластеров) $\mathbb{C}_1, ..., \mathbb{C}_K$, каждый из которых лучше всего аппроксимируется его d-мерным аффинным подпространством $\mathbb{F}_k$, $k = \overline{1, K}$. При заданных K и d этот алгоритм минимизирует целевую функцию

$$e = \sum_{k=1}^{K} \min_{\mathbb{F}_k,\, k=\overline{1,K}} \sum_{\mathbf{z}_j \in \mathbb{C}_k} \|\mathbf{z}_j - P_{\mathbb{F}_k}\mathbf{z}_j\|^2,$$

где $P_{\mathbb{F}_k}\mathbf{z}_j$ — проекция вектора $\mathbf{z}_j$ на d-мерное аффинное подпространство $\mathbb{F}_k$, $k = \overline{1, K}$. На практике минимизация целевой функции выполняется итеративно, как в алгоритме K-средних [8]. То есть после инициализации K d-мерных аффинных подпространств (например, они могут быть выбраны случайным образом) повторяются два шага до достижения сходимости: 1) назначаются кластеры в соответствии с минимальным расстоянием до аффинных подпространств, определенных на предыдущей итерации; 2) для этих вновь полученных кластеров посредством анализа главных компонент — *Principal Component Analysis* (PCA [9]) вычисляются d-мерные аффинные подпространства с минимальной средней квадратичной ошибкой. Эта процедура очень быстрая и гарантированно сходится, по крайней мере, к локальному минимуму. Однако на практике локальный минимум, к которому сходится алгоритм K аффинных подпространств, гораздо хуже глобального минимума целевой функции. В результате, этот алгоритм не такой точный, как более ранние алгоритмы гибридного линейного моделирования, и даже при моделировании поверх линейных подпространств (в противоположность общим аффинным подпространствам) он часто дает сбой, когда, или d достаточно велико (например $d \geqslant 10$), или имеется значительная составляющая посторонних векторов [10, 11].

Ниже предполагается, что набор векторов $\mathbb{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\} \subset \mathbb{R}^D$ выбран из объединения K d-мерных аффинных подпространств $\mathbb{F}_k$, $k = \overline{1, K}$ (возможно с шумом и посторонними векторами), где $K > 1$ и N велико. Используя полярную кривизну c_{p} (2) с фиксированной константой моделирования $\sigma > 0$, конструируется тензор аффинности

(близости) $\mathcal{A}$ порядка $(d + 2)$ для различных $(d+2)$ векторов $\mathbf{z}_{i_1}, ..., \mathbf{z}_{i_{d+2}} \in \mathbb{Z}$ с компонентами

$$\mathcal{A}(i_1, ..., i_{d+2}) = \exp(-(c_{\mathrm{p}}(\mathbf{z}_{i_1}, ..., \mathbf{z}_{i_{d+2}}))^2/(2\sigma^2)). \quad (3)$$

Выбор оптимального значения параметра σ обсуждается в работе [2]. В выражении (3) тензор близости (подобия) $\mathcal{A}$ порядка $(d + 2)$ имеет размерность $N \times N \times ... \times N$. Однако в настоящей работе используется только матричное представление тензора $\mathcal{A}$ (3), которое обозначается **A** и называется матрицей близости (подобия). Размерность матрицы **A** равна $N \times N^{d+1}$. Для каждого $i = \overline{1, N}$ строка i матрицы **A** (т. е. **A**$(i, :)$) развертывается из слоя i тензора $\mathcal{A}$ (3) (т. е. $\mathcal{A}(i, :, ..., :)$), следуя некоторому произвольному, но фиксированному порядку, например лексикографическому порядку последних $(d + 1)$ индексов [12]. Это упорядочение несущественно для настоящего рассмотрения, поскольку как показано в работах [4, 5], в спектральной кластеризации используется произведение

$$\mathbf{W} = \mathbf{A}\mathbf{A}^{\mathrm{T}}, \quad (4)$$

которое не зависит от порядка индексов. Определение матрицы **A** и умножение этой матрицы большой размерности на результат ее транспонирования (чтобы вычислить **W** (4)) выполнить сложно. Возможное решение состоит в однородной выборке, т. е. в случайной выборке и вычислении небольшого поднабора столбцов **A**, чтобы провести оценку **W** (4) [3, 13]. Обозначая посредством **A**$(:, j)$ столбец j матрицы **A**, записываем **W** (4) в следующем виде:

$$\mathbf{W} = \sum_{j=1}^{N^{d+1}} \mathbf{A}(:, j)\mathbf{A}(:, j)^{\mathrm{T}}. \quad (5)$$

Следовательно, **W** — сумма N^{d+1} матриц ранга 1. Пусть $j_1, ..., j_c$ — c целых чисел, случайно выбранных из интервала $[1, N^{d+1}]$. Как показано в работе [13], матрица **W** (5) аппроксимируется следующим образом:

$$\mathbf{W} \approx \sum_{t=1}^{c} \mathbf{A}(:, j_t)\mathbf{A}(:, j_t)^{\mathrm{T}}. \quad (6)$$

В работе [2] алгоритм спектральной кластеризации с матрицей близости на основе полярной кривизны (2) — *Spectral Curvature Clustering* (SCC) формирует матрицу парных весов **W** (6) из аппроксимированной матрицы близости (подобия) $\mathbf{W} = \mathbf{A}_c\mathbf{A}_c^{\mathrm{T}}$, и применяет спектральную кластеризацию [4] для определения K кластеров $\mathbb{C}_1, ..., \mathbb{C}_K$. Для того чтобы улучшить эти кластеры, алгоритм SCC затем повторно выбирает векторы из кластеров $\mathbb{C}_k$, $k = \overline{1, K}$, в пределах небольшой полосы вокруг каждого из их аппроксимирующих d-мерных аффинных подпространств $\mathbb{F}_k$, $k = \overline{1, K}$. Эта процедура повторяется до достижения сходимости и называется итеративной выборкой (*iterative sampling*) [2]. Сходимость измеряется ошибкой ортогональных наименьших квадратов — *Orthogonal Least Squares* (OLS) для d-мерных аффинных подпространств $\mathbb{F}_k$, $k = \overline{1, K}$, аппроксимирующих кластеры $\mathbb{C}_1, ..., \mathbb{C}_K$ в виде

$$e_{\mathrm{OLS}} = \sum_{k=1}^{K} \sum_{z_j \in \mathbb{C}_k} \|\mathbf{z}_j - P_{\mathbb{F}_k}\mathbf{z}_j\|^2, \quad (7)$$

где $P_{\mathbb{F}_k}\mathbf{z}_j$ — проекция вектора $\mathbf{z}_j$ на d-мерное аффинное подпространство $\mathbb{F}_k$, $k = \overline{1, K}$ (может быть получена анализом главных компонент (PCA) [8]).

Алгоритм спектральной кластеризации с матрицей близости (подобия) SCC на основе полярной кривизны представлен ниже [2].

Вход: Набор векторов $\mathbb{Z}$, внутренняя размерность (*intrinsic dimension*) d, число K d-мерных аффинных подпространств, число выбираемых столбцов c (по умолчанию $= 100K$).

Выход: K непересекающихся кластеров $\mathbb{C}_1, ..., \mathbb{C}_K$ и ошибка e_{OLS}.

Шаги:

1. Случайным образом выбираются c поднаборов из $\mathbb{Z}$, каждый из которых содержит в точности $(d + 1)$ различных векторов.

2. Вычисляются полярная кривизна (2) каждого поднабора и каждого из оставшихся $(N - d - 1)$ векторов в $\mathbb{Z}$, эти $(N - d - 1)c$ значений полярной кривизны сортируются по возрастанию и формируют вектор $\mathbf{c}_{\mathrm{p}}$.

3. **for** $q = 1$: $(d + 1)$**do**

- Используется (3) с $\sigma = \mathbf{c}_{\mathrm{p}}(Nc/K^q)$ для вычисления c выбранных столбцов **A**. Используя эти c столбцов, формируется матрица $\mathbf{A}_c \in \mathbb{R}^{N \times c}$ (6).
- Вычисляется матрица $\mathbf{D} = \mathrm{diag}(\mathbf{A}_c(\mathbf{A}_c^{\mathrm{T}}\mathbf{1}))$, где **1** — вектор из единиц, и эта матрица используется для нормировки матрицы $\mathbf{A}_c$: $\tilde{\mathbf{A}}_c = \mathbf{D}^{-1/2}\mathbf{A}_c$.
- Формируется матрица **U**, столбцы которой — K старших левых сингулярных векторов $\tilde{\mathbf{A}}_c$.
- К строкам **U** применяется алгоритм K-средних (возможно, нормированным на единичную длину), и они разделяются на K кластеров.
- Эти кластеры (найденные) используются для группировки векторов набора $\mathbb{Z}$ в K поднаборов, и вычисляется соответствующая ошибка e_{OLS} (7).

end for

Регистрируются K поднаборов $\mathbb{C}_1, ..., \mathbb{C}_K$ набора $\mathbb{Z}$, которые соответствуют наименьшей ошибке e_{OLS} (7) в приведенном выше цикле.

4. Из каждого найденного $\mathbb{C}_K$, $k = \overline{1, K}$, выбираются вектора в пределах небольшой полосы вокруг каждого из их OLS-аппроксимирующих d-мерных аффинных подпространств $\mathbb{F}_k$, $k = \overline{1, K}$, и выполняются шаги 2 и 3, чтобы найти K новых кластеров. Итерации повторяются до достижения сходимости.

2. Численный эксперимент

В настоящей работе анализ тонкой структуры кластеров траекторий движения самолетов, полученных методом полиномиальных регрессий [14], выполняется с помощью алгоритма спектральной кластеризации с тензором близости (подобия) на основе полярной кривизны (2).

Используются траектории 117 самолетов, идущих на посадку в международном аэропорту и зарегистрированных радаром TRACON 1 января 2006 года (https://c3.nasa.gov/dashlink/resources/132/). Начало координат совпадает с положением радара, интервал времени между точками регистрации составляет около 5 с. В работе учитываются только 160 последних точек каждой траектории, что позволяет исключить случайные маневры самолетов перед заходом на посадку. Эти траектории в трехмерном пространстве представлены в работе [14]. Пять кластеров траекторий самолетов (рис. 1, см. третью сторону обложки), выделяются в результате применения метода полиномиальных регрессий. Каждый кластер соответствует определенному "посадочному" паттерну или маршруту. Распределение числа траекторий по кластерам следующее: 16 траекторий в розовом кластере; 13 — в зеленом; 3 — в синем; 37 — в черном и 38 — в красном кластерах.

На рис. 2 (см. третью сторону обложки) показаны проекции траекторий в анализируемых кластерах (см. рис. 1) на координатные оси x, y и z в последовательные моменты времени k регистрации радаром. Линии тренда по методу полиномиальных регрессий (выделены жирными линиями) представляют обобщенную форму траекторий в каждом кластере [14]. Сходство траекторий движения самолетов в кластерах отражает несколько типичных маршрутов посадки ("посадочных" паттернов).

В работах [15, 16] по модели линейных и нелинейных динамических систем в четырех из пяти рассматриваемых кластеров при рассмотрении проекций траекторий на координатные оси x, y, z выявляется их тонкая (неоднородная) структура.

Восстановление двух первых столбцов исходных данных после разложения сингулярных чисел (svd-разложения) и сокращение числа сингулярных чисел до двух показано на рис. 3 (в коде MATLAB исходные данные — переменная x).

На рис. 4 (см. третью сторону обложки) для первичного красного кластера (см. рис. 1 и рис. 2) для x-компоненты траекторий при внутренней размерности аффинных подпространств $d = 5$ показаны три новых подкластера: красный, синий и зеленый. На рис. 4, *а* показана кластеризация алгоритмом SCC в координатах: $data_1$ = `data(:, 1)` и $data_2$ = `data(:, 2)` (см. рис. 3). На рис. 4, *в* алгоритм SCC в общей модели 5-мерных аффинных подпространств ошибочно относит розовую траекторию к красному подкластеру, и в этом подкластере по мере косинуса она оказывается кандидатом на постороннюю. В результате нормальная траектория зеленого подкластера идентифицируется как посторонняя (показана желтым цветом на рис. 4, *в*). В синем подкластере голубым цветом показан кандидат на постороннюю траекторию. Классификация алгоритмом LSCC, предложенным в работе [2] (см. рис. 4, *б*), показана в координатах двух первых столбцов восстановленных данных (см. рис. 3). На рис. 4, *г* алгоритм LSCC в модели 5-мерных линейных подпространств правильно относит эту траекторию к зеленому подкластеру, но в этом подкластере она оказывается кандидатом на постороннюю (показана желтым цветом). Теперь в красном подкластере по мере косинуса правильно определяется кандидат на постороннюю траекторию (показана розовым цветом).

```
[N,D] = size(X);
[U,S] = svds(X - repmat(mean(X,1),N,1),6);
data = U(:,1:2).*repmat(transpose(diag(S(1:2,1:2))),
N,1);
```

Рис. 3. Восстановление исходных данных после разложения сингулярных чисел и сокращение их числа до двух

Заключение

В работе показано, как кластеризация траекторий движения самолетов при посадке в аэропорту осуществляется геометрическим методом, при котором набор многомерных векторов моделируется объединением d-мерных аффинных подпространств, заменяемых центроидами как прототипами кластеров. Поэтому учитывается не бинарное, а d-арное подобие, и принадлежность рассматриваемого набора траекторий (многомерных векторов) к кластеру оценивается тензором (многомерным) близости порядка $(d + 2)$, который разворачивается в матрицу близости (подобия) векторов, обеспечивая эффективное использование спектрального метода. Оценка близости в кластере (по мере косинуса) обнаруживает потенциально посторонние траектории движения самолетов в зоне риска.

Список литературы

1. **Ma Y., Yang A. Y., Derksen H., Fossum R.** Estimation of subspace arrangements with applications in modeling and segmenting mixed data // SIAM Review. 2008. Vol. 50, N. 3. P. 413—458.
2. **Chen G., Lerman G.** Spectral curvature clustering (SCC) // International Journal on Computer Vision. 2009. Vol. 81, N. 3. P. 317—330.
3. **Govindu V.** A tensor decomposition for geometric grouping and segmentation // Proc. of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). 20—26 June 2005, San Diego, CA. Washington, DC; IEEE Computer Society. 2005. V. 1. P. 1150—1157.
4. **Ng A., Jordan M., Weiss Y.** On spectral clustering: Analysis and an algorithm / Eds. Dietterich T., Becker S. Ghahramani Z. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press. 2002. V. 14. P. 1—8.
5. **Luxburg U.** A tutorial on spectral clustering // Statistics and Computing. December 2007. Vol. 17, Is. 4. P. 395—416.
6. **Daverman R. J., Sher R. B.** Handbook on Geometric Topology. Amsterdam: Elsevier Science BV, 2002, 1133 p.

7. **Whitehouse J. T.** Generalized Sines, Multiway Curvatures, and the Multiscale Geometry of d-Regular Measures. PhD Dissertation. University of Minnesota: Faculty of the Graduate School, 2009.
8. **Kroonenberg P. M.** Applied Multiway Data Analysis. Hoboken, NJ: Wiley, 2008. 557 p.
9. **Jolliffe I. T.** Principal Component Analysis. New York, Berlin, Heidelberg: Springer, 2002. 406 p.
10. **Zhang T., Szlam A., Lerman G.** Median *K*-flats for hybrid linear modeling with many outliers // Proc. of the 2009 IEEE 12th International Conference on Computer Vision, Sept. 27 — Oct. 4, 2009. Washington, DC: IEEE Computer Society 2009. P. 234—241.
11. **Zhang T., Szlam A., Wang Y., Lerman G.** Hybrid linear modeling via local best-fit flats // International Journal of Computer Vision. 2012. Vol. 100, N. 3. P. 217—240.
12. **Bader B., Kolda T.** Matlab Tensor Classes for Fast Algorithm Prototyping. Technical Report SAND2004-5187. Albuquerque, NM: Sandia National Laboratories, 2004.
13. **Drineas P., Kannan R., Mahoney M.** Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication // SIAM Journal on Computing. 2006. Vol. 36, N. 1. P. 132—157.
14. **Кухаренко Б. Г., Солнцева М. О.** Кластеризация управляемых объектов на основе сходства их многомерных траекторий // Информационные технологии. 2014. № 5. С. 3—7.
15. **Кухаренко Б. Г., Солнцева М. О.** Анализ результатов кластеризации многомерных траекторий посредством моделей линейных динамических систем // Информационные технологии. 2015. Т. 21, № 2. С. 104—109.
16. **Кухаренко Б. Г., Солнцева М. О.** Применение моделей нелинейных динамических систем для анализа результатов кластеризации многомерных траекторий // Информационные технологии. 2015. Т. 21, № 5. С. 341—345.

B. G. Kukharenko, e-mail: kukharenkobg@hotmail.com,
Leading Research Scientist, Blagonravov Mechanical Engineering Research Institute of the RAS,
M. O. Solntseva-Chalei, Graduate Student, e-mail: solnceva.chalei@gmail.com,
Moscow Institute of Physics and Technology (SU)

Spectral Method with Polar Curvature Used for Analysis of Multi-Dimensional Trajectory Clustering Results

Problem of clustering trajectories is pre-conditioned by a need to organize motion of objects under control. The polynomial regression method is the best approach to trajectory cluster selection, which estimates a form of general trajectory of each cluster. For a set of sufficiently heterogeneous trajectories, the defined clusters are heterogeneous also. For the inhomogeneous cluster, its polynomial regression is a too strong abstraction. To demonstrate the cluster heterogeneity (and, thus, non full clustering trajectories) a method of dimension reducing is in need. As K d-flats algorithm is a generalization of K-means algorithm where d-dimensional best fit affine sets replace centroids as the cluster prototypes, a similar modification of spectral clustering method is in use in paper. This is geometric method to obtain vector clustering model in form of affine space union. Vector set similarity is estimated by multi-dimensional affinity tensor, for which matrix representation is in use. Next it is analyzed by spectral method. As example, clustering results of airplane fly trajectories in an airport space is under study. Defining the cluster central trajectory with affinity estimation by cosine measure in use gives an opportunity to determine the cluster outlying trajectories in potential, which represent fly trajectories located in a risk zone.

Keywords: *data mining, multi-dimensional trajectories, model of affine space union, polar curvature, spectral clustering*

References

1. **Ma Y., Yang A. Y., Derksen H., Fossum R.** Estimation of subspace arrangements with applications in modeling and segmenting mixed data, *SIAM Review*, 2008, vol. 50, no. 3, pp. 413—458.
2. **Chen G., Lerman G.** Spectral curvature clustering (SCC), *International Journal on Computer Vision*, 2009, vol. 81, no. 3, pp. 317—330.
3. **Govindu V.** A tensor decomposition for geometric grouping and segmentation, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 20—26 June 2005, San Diego, CA, Washington, DC: IEEE Computer Society, 2005, vol. 1, pp. 1150—1157.
4. **Ng A., Jordan M., Weiss Y.** On spectral clustering: Analysis and an algorithm / Dietterich T., Becker S. Ghahramani Z., eds. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press. 2002, vol. 14, pp. 1—8.
5. **Luxburg U.** A tutorial on spectral clustering, *Statistics and Computing*, December 2007, vol. 17, is. 4, pp. 395—416.
6. **Daverman R. J., Sher R. B.** Handbook on Geometric Topology. Amsterdam: Elsevier Science BV, 2002, 1133 p.
7. **Whitehouse J. T.** *Generalized Sines, Multiway Curvatures, and the Multiscale Geometry of d-Regular Measures.* PhD Dissertation. University of Minnesota: Faculty of the Graduate School, 2009.
8. **Kroonenberg P. M.** *Applied Multiway Data Analysis.* Hoboken, NJ: Wiley, 2008. 557 p.
9. **Jolliffe I. T.** Principal Component Analysis. New York, Berlin, Heidelberg: Springer, 2002, 406 p.
10. **Zhang T., Szlam A., Lerman G.** Median K-flats for hybrid linear modeling with many outliers, *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, Sept. 27 — Oct. 4, 2009. Washington, DC: IEEE Computer Society, 2009, pp. 234—241.
11. **Zhang T., Szlam A., Wang Y., Lerman G.** Hybrid linear modeling via local best-fit flats, *International Journal of Computer Vision*, 2012, vol. 100, no. 3. P. 217—240.
12. **Bader B., Kolda T.** Matlab Tensor Classes for Fast Algorithm Prototyping. Technical Report SAND2004-5187. Albuquerque, NM: Sandia National Laboratories, 2004.
13. **Drineas P., Kannan R., Mahoney M.** Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication, *SIAM Journal on Computing*, 2006, vol. 36, no. 1, pp. 132—157.
14. **Kukharenko B. G., Solntseva M. O.** Klasterizacia upravlyaemyh objektov na osnove shodstva ih mnogomemyh trajektoryi, *Informacionnye tehnologii*, 2014, no. 5, pp. 3—7.
15. **Kukharenko B. G., Solntseva M. O.** Analiz resultatov klasterizacii mnogomernyh trajektoryi posredstvom modelei lineinyh dinamicheskih sistem, *Informacionnye tehnologii*, 2015, no. 2, pp. 104—109.
16. **Kukharenko B. G., Solntseva M. O.** Primenenie modelei nelineinyh dinamicheskih sistem dlya analiza resultatov klasterizacii mnogomernyh trajektoryi, *Informacionnye tehnologii*, 2015, vol. 21, no. 5, pp. 341—345.