Исследование периодичностей в пространственно-временных данных

Векличева Мария Константиновна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., асс. Коробейников А.И.
Рецензент: к.ф.-м.н., доцент Голяндина Н.Э.

СтатМод

Санкт-Петербург
2012г.

Пример

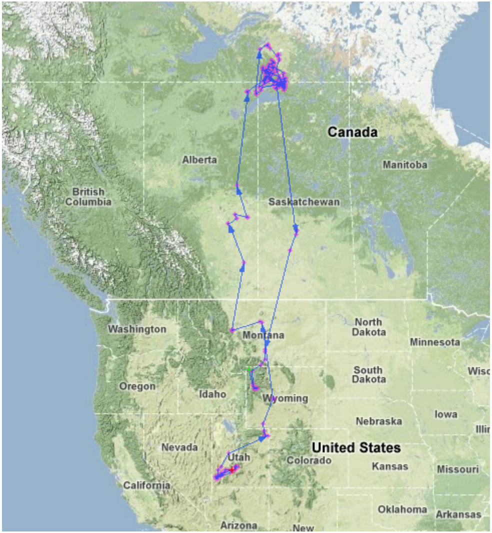

Рис.: Путь миграции орла.

Базовый алгоритм: Регуляризация и Кластеризация пространственных данных

Алгоритм был предложен в статье Z. Li et al.(2010) и обширно используется ресурсом `http://movebank.org/`.

Регуляризация данных:

$$D^0 = \{(x_j^0, y_j^0, t_j^0)\}_{j=1}^N \mapsto D = \{(x_i, y_i, t_i)\}_{i=1}^n,$$

где (x_i, y_i) — географические координаты в момент времени t_i и $t_2 - t_1 = t_3 - t_2 = \ldots = t_n - t_{n-1}$.

Кластеризация пространственных данных:

$$(x_i, y_i) \mapsto c_i,$$

где $c_i \in \{\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_M\}$ и $c_i = \mathcal{O}_0$, если (x_i, y_i) — выброс.

Базовый алгоритм: Определение значений периодов

$\{c_i\}_{i=1}^{n}$ — последовательность всех индексов кластеров

Определение

Период T характерен для кластера $\mathcal{O}_j$, если найдется i :

$$c_i = c_{i+T} = c_{i+2T} = \ldots = c_{i+kT} = \mathcal{O}_j,$$

где $1 \leqslant i \leqslant T$ и $k \in \mathbb{N} : i + kT \leqslant n$. Равенство может не выполняться лишь для некоторого небольшого числа случаев.

Для каждого $\mathcal{O}_j$:

1. $\mathcal{O}_j \mapsto B^j = b_1^j \ldots b_n^j$, где
$$b_i^j = \begin{cases} 1, & \text{если } c_i = \mathcal{O}_j, \\ 0, & \text{иначе.} \end{cases}$$
2. $B^j \mapsto I^j(\lambda)$ — периодограмма;
3. $I^j(\lambda) \mapsto \{T_1^j, \ldots, T_l^j\}$ — периоды.

Для каждого периода T:
$\{\mathcal{O}_{s_1}, \ldots, \mathcal{O}_{s_d}\}$ — кластеры с характерным T
$\mathcal{O}_{s_0}$ — все остальные

$$c_1 c_2 c_3 \ldots c_n \mapsto I^1, \ldots, I^m \text{ – сегменты :}$$

$$\underbrace{\tilde{c}_1 \tilde{c}_2 \ldots \tilde{c}_T}_{I^1} \underbrace{\tilde{c}_{T+1} \tilde{c}_{T+2} \ldots \tilde{c}_{2T}}_{I^2} \ldots \underbrace{\tilde{c}_{(m-1)T+1} \tilde{c}_{(m-1)T+2} \ldots \tilde{c}_{mT}}_{I^m},$$

где $m = \lfloor \frac{n}{T} \rfloor$ и

$$\tilde{c}_i = \begin{cases} c_i, & \text{если } c_i \in \{\mathcal{O}_{s_1}, \ldots, \mathcal{O}_{s_d}\}, \\ \mathcal{O}_{s_0}, & \text{иначе.} \end{cases}$$

Определение (Модель периодического поведения для T)

Периодическое поведение μ — это многомерная случайная величина вида $(\mu_1, \ldots, \mu_T)$, где все μ_i независимы и μ_i принимает значения из $\{\mathcal{O}_{s_0}, \mathcal{O}_{s_1}, \ldots, \mathcal{O}_{s_d}\}$.

Базовый алгоритм: Оценка распределения

Распределение случайной величины $\mu = (\mu_1, \ldots, \mu_T)$ однозначно определяется матрицей вида $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_T]$, где $\mathbf{p}_k$ задает распределение величины μ_k.

Вероятность того, что произвольный набор сегментов $\mathcal{I} = \bigcup_{j=1}^{l} I^j$ порожден $\mu \sim \mathbf{P}$:

$$\mathsf{P}(\mathcal{I} \mid \mathbf{P}) = \prod_{I^j \in \mathcal{I}} \prod_{k=1}^{T} \mathsf{P}(\mu_k = I_k^j).$$

Оценка матрицы $\mathbf{P}$:

$$\hat{p}_{ik} = \frac{\sum_{I^j \in \mathcal{I}} \mathbf{1}_{I_k^j = i}}{|\,\mathcal{I}\,|}$$

Базовый алгоритм: Кластеризация распределений

Задача:
Из m распределений, полученных как оценки распределений по каждому сегменту I^j, выделить ρ распределений, где $\rho \leqslant m$ и заранее неизвестно.

Решение: Выполнить иерархическую кластеризацию с использованием расстояния Кульбака-Лейблера, которое для двух распределений, задаваемых матрицами $\mathbf{P}$ и $\mathbf{Q}$, имеет вид:

$$KL(\mathbf{P}, \mathbf{Q}) = \sum_{k=1}^{T} \sum_{i=0}^{d} p_{ik} \log \frac{p_{ik}}{q_{ik}}.$$

Причины модификаций

- Регуляризация данных до пространственной кластеризации.
- Вычислительная сложность пространственной кластеризации.
- Оценка распределения по одному сегменту.

Метод:
Иерархическая агломеративная кластеризация с расстоянием на Земном шаре.

Оптимальное деление на кластеры:
Метод динамического подрезания дендрограммы (P. Langfelder et al., 2008).

- Метод основывается на анализе формы ветвей дендрограммы, являющейся результатом иерархической кластеризации, а также матрице расстояний между кластеризуемыми объектами.

Модификации: Определение наличия периодичности

$I(\omega_j)$ — периодограмма в точках $\{\omega_j = 2\pi j/l\}_{j=1}^{q}$ для последовательности X длины l, $q = \lfloor (l-1)/2 \rfloor$.

1. Перестановочная процедура.

2. Точный g-тест Фишера:

$$\xi_q = \frac{\max_{1\leqslant j\leqslant q} I(\omega_j)}{q^{-1}\sum_{i=j}^{q} I(\omega_j)}.$$

3. Критерий распределения максимума периодограммы (R. A. Davis, T. Mikosch, 1999):

$$M_l(X) = \max_{1\leqslant j\leqslant q} I(\omega_j).$$

Симметричные расстояния:

1. расстояние Йенсона-Шеннона:

$$JSD^2(\mathbf{P},\mathbf{Q}) = \sum_{i=0}^{d}\sum_{k=1}^{T}\left(p_{ik}\log\frac{p_{ik}}{r_{ik}} + q_{ik}\log\frac{q_{ik}}{r_{ik}}\right),\ \mathbf{R} = \frac{1}{2}(\mathbf{P}+\mathbf{Q}).$$

2. расстояние Хеллинджера:

$$He^2(\mathbf{P},\mathbf{Q}) = \sum_{i=0}^{d}\sum_{k=1}^{T}\left(\sqrt{p_{ik}} - \sqrt{q_{ik}}\right)^2.$$

3. условное:

$$C_D(\mathbf{P},\mathbf{Q}) = D(\hat{\mathbf{P}},\hat{\mathbf{Q}}),$$

где D — расстояние между $\mathbf{P}$ и $\mathbf{Q}$, а $\hat{\mathbf{A}}$ означает матрицу размера $d \times T$, полученную из матрицы $\mathbf{A}$ и имеющую вид:

$$[\hat{\mathbf{a}}_1,\ldots,\hat{\mathbf{a}}_T], \hat{\mathbf{a}}_k = (a_{1k},\ldots,a_{dk})^T / \sum_{i=1}^{d} a_{ik}, k = 1,\ldots,T.$$

Модификации: Кластеризация распределений

Выполнить промежуточную кластеризацию распределений:
Для m дискретных T-мерных распределений $\{\mathbf{p}^1, \ldots, \mathbf{p}^m\}$ найти ρ распределений $\{\mathbf{c}^1, \ldots, \mathbf{c}^\rho\}$ из их выпуклой оболочки Δ таких, чтобы:

$$\{\mathbf{c}^1, \ldots, \mathbf{c}^\rho\} = \underset{\mathbf{c}^{i_1}, \ldots, \mathbf{c}^{i_\rho} \in \Delta}{\arg\min} \left(\sum_{j=1}^{m} \min_{1 \leqslant s \leqslant \rho} KL(\mathbf{p}^j, \mathbf{c}^{i_s}) \right)$$

Решение (K. Chaudhuri et al., 2008 год):
Решить задачу кластеризации распределений методом k-средних с использованием расстояния Хеллинджера.

Моделирование данных

Параметры:

- доля шумовых непериодических наблюдений для одного поведения (от 0 до 0.5 с шагом 0.1);
- дисперсия шума, добавляемого к временным меткам (от 0 до 2 часов с шагом в 15 минут);
- количество известных наблюдений в течение периода (от 2 до 24 с шагом 2).

Моделирование данных:

1. моделирование распределений для каждого поведения;
2. моделирование последовательности перемещений и временных меток;
3. моделирование географических координат.

Для простоты последующего анализа результатов моделируются данные с единственным периодом в 24 часа и двумя поведениями.

Вариации алгоритма

1. Определение периодичности:
   - перестановочная процедура;
   - точный g-тест Фишера;
   - критерий распределения максимума периодограммы.
2. Нахождение периодических поведений:
   - методы:
     - иерархическая кластеризация;
     - сочетание предварительного метода k-средних Чаудхури и иерархической кластеризации.
   - расстояния:
     - расстояние Кульбака-Лейблера;
     - расстояние Хеллинджера;
     - расстояние Йенсона-Шеннона;
     - условные.
   - метод k-средних Чаудхури.

Результаты моделирования

Ошибка между начальным поведением с распределением, задаваемым матрицей $\mathbf{P}$, и полученным в результате применения алгоритма с матрицей $\mathbf{Q}$ вычисляется:

$$\frac{1}{(d+1)\cdot T}\sum_{i=0}^{d}\sum_{k=1}^{T}|p_{ik}-q_{ik}|.$$

Основные выводы:

- Применение предварительной кластеризации в большинстве случаев приводит к улучшению результатов.
- Использование расстояний Хеллинджера и Йенсона-Шеннона приводит к схожим результатам, которые лучше результатов, полученных с использованием расстояния Кульбака-Лейблера.
- Условные расстояния не дают лучших результатов, чем традиционные расстояния.

Обнаруженные периоды:

- 24 часа
- 168 часов.

Выявили:

- 24 часа: 2 поведения.
- 168 часов: 1 поведение.

Итоги

- Предложены модификации базового алгоритма.
- Исследовано качество работы базового алгоритма и алгоритма с модификациями на модельных данных разного качества.
- Реализованы в математической среде R базовый алгоритм, модификации и моделирование.
- Рассмотрен пример выполнения алгоритма на реальных данных.