Статистические методы кластеризации в больших объемах данных

Зиннатулина Белла Раифовна, гр. 622

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Научный руководитель: д. ф.-м. н., профессор М.С. Ермаков
Рецензент: д. ф.-м. н., профессор Г.Л. Шевляков

Санкт-Петербург
2017г.

Мотивация

- **Источники больших данных:** социальные сети, устройства видеорегистрации, метеорологические данные и т.д.
- **Проблемы:**
  - Необходимо хранение и своевременная обработка.
  - Основной объем данных — неструктурированная, непрерывно поступающая информация.
  - Разреженная структура данных.
  - Наличие как качественных, так и количественных признаков.
- **Специфика больших данных:** объем, скорость, многообразие, ценность
- Цели кластерного анализа:
  - Понимание данных путём выявления кластерной структуры
  - Редукция размерности данных
  - Обнаружение новизны

Мотивация

- **Источники больших данных:** социальные сети, устройства видеорегистрации, метеорологические данные и т.д.
- **Проблемы:**
  - Необходимо хранение и своевременная обработка.
  - Основной объем данных — неструктурированная, непрерывно поступающая информация.
  - Разреженная структура данных.
  - Наличие как качественных, так и количественных признаков.
- **Специфика больших данных:** объем, скорость, многообразие, ценность
- **Цели кластерного анализа:**
  - Понимание данных путём выявления кластерной структуры
  - Редукция размерности данных
  - Обнаружение новизны

Цель и задачи

Цель

Изучение статистических методов кластеризации больших объемов данных с их практическим применением.

В **задачи** работы входило:

- Изучение специфики больших данных;
- Анализ литературы на предмет современных методов кластеризации;
- Выбор наиболее интересного для изучения метода;
- Реализация выбранного алгоритма;

Обзор методов

- Группа методов иерархической кластеризации согласно [Lance G.N., Williams W.T.];
  - Строят систему вложенных разбиений
  - \+ Разнообразние мер подсчета расстояния между точками пространства
  - \- Сложности при работе с категориальными данными
- Алгоритм кластеризации Маркова [Stijn van Dongen];
  - Основан на моделировании случайных блужданий в графах
  - \+ Идея проста в реализации: поочередное применение двух операторов
  - \- Необходимо распараллеливание, т.к. основан на перемножении больших матриц
- Стохастическая блочная модель [Emmanuel Abbe]

Стохастическая блочная модель (SBM)

- Пусть V — множество вершин графа размера $n \in \mathbb{N}$.
- $k \in \mathbb{N}$ — количество кластеров.
- $p = (p_1, \dots, p_k)$ — вектор вероятностей принадлежности объекта кластеру.

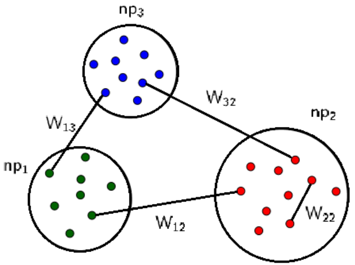

Рис.: Пример при $k = 3$.

- Пусть $X = (X_1, \dots, X_n)$, $X_i \in \{1, \dots, k\}$ с вероятностью p_i. Тогда $X_i \in X$, $i \in 1, \dots, n$ — метка вершины $V_i \in V$.
- Пусть W — симметричная матрица размера $k \times k$.
- $W_{ij} \in [0, 1]$ — вероятности связи вершин графа.

  Пара вершин $(V_i, V_j) \in V \times V$, $i, j \in 1, \cdots, n$ связаны ребром с вероятностью $W_{X_i X_j} \in W$.

Стохастическая блочная модель (SBM)

Цель обнаружения кластера

Восстановление разметки X с некоторым уровнем точности путем наблюдения G. Результат – k кластеров размера $np_1, np_2, \dots, np_k$.

Пара (X, G) строится с помощью SBM(n, p, W), где

- G — неориентированный граф из n вершин
- вершины i и j соединены ребром с вероятностью $W_{X_i X_j}$ и независимо от других пар вершин.

Условия восстановимости [Abbe, 2015]

Пусть $\hat{X}$ — любое преобразование элементов вектора X фиксированной перестановки $[k]$.

Согласие α между векторами X и $\hat{X}$ получается путем минимизации расстояния Хэмминга между ними.

- **Точное** восстановление разрешимо в SBM(n, p, W), если существует алгоритм с точностью $\alpha = 1$.
- **Сильное** восстановление разрешимо в SBM(n, p, W), если существует алгоритм с точностью $\alpha = 1 - o_n(1)$.
- **Слабое** восстановление разрешимо в SBM$(n, u, \hat{W})$, (где u — равномерно распределение на $[k]$, а $\hat{W}$ — матрица, с константой вне диагонали) если существует алгоритм с точностью $\alpha = \frac{1}{k} + \varepsilon, \varepsilon > 0$.

Условие точной восстановимости

Пусть $W = S_n \frac{Q}{n}$, где $Q \in \mathrm{R}_+^{k\times k}$ — симметричная матрица, не зависящая от n, S_n — параметр интенсивности, определяется количеством вершин графа.

Определим пороговое значение SNR(Kesten-Stigum threshold) следующим образом:

$$\mathrm{SNR} = \frac{|\lambda_{min}|^2}{|\lambda_{max}|},$$

где $|\lambda_{min}|$ и $|\lambda_{max}|$ — наименьшее и наибольшее собственное число матрицы $\mathrm{diag}(p)Q$ соответственно.

Теорема (Emmanuel Abbe, 2016)

Точное восстановление разрешимо в SBM$(n, p, \log(n)Q/n) \iff J(p, Q) := \min_{1\le i<j\le k} D_+((\mathrm{diag}(p)Q)_i || (\mathrm{diag}(p)Q)_j) \geq 1$, *где* D_+ *определено как*

$$D_+(\mu||\nu) = \max_{t\in[0,1]} \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) = 1 - t + ty - y^t.$$

Симметричная стохастическая блочная модель (Symmetric SBM)

- Случай равновероятной кластеризации, где $p = (\frac{1}{k}, \cdots, \frac{1}{k})$;
- Матрица W имеет следующий вид:

$$W = \begin{pmatrix} a & b & \dots & b \\ b & \ddots & b & \vdots \\ \vdots & b & \ddots & b \\ b & \dots & b & a \end{pmatrix}$$

- В данной модели нет разницы между группами, поэтому вероятности связи между всеми кластерами равны между собой, так же как и вероятности связи внутри кластеров.

Симметричная стохастическая блочная модель (Symmetric SBM)

Обозначим через SSBM$(n, k, \frac{a}{n}, \frac{b}{n})$ симметричный разреженный SBM(n, p, W), где $p \in \mathrm{U}(1, k)$ и

$$W_{ij} = \begin{cases} a/n, & \text{если } i = j, \\ b/n, & \text{иначе.} \end{cases}$$

Определим SNR в случае k симметричных кластеров:

$$\mathrm{SNR} = \frac{(a-b)^2}{k(a+(k-1)b)},$$

тогда

- не зависимо от k, если SNR > 1, то задача распознавания кластеров разрешима за полиномиальное время;
- если $k \geq 5$, то задача разрешима для некоторого SNR < 1.

1. **Вход**: пара (G, γ), где G — исходный граф, $\gamma \in [0, 1]$.
2. Определим подграф G' на множестве вершин $1 \dots n$, где каждое ребро из графа G выбирается независимо с вероятностью γ.
3. Считаем $I_{r,r'[E]}(v_i \cdot v_j)$ для всех i и j графа G'.
4. Существует такое разбиение вершин i и j, что $I_{r,r'[E]}(v_i \cdot v_j) > 0$ $\iff$ когда вершины i и j лежат в одном кластере. Выбираем по одному из представителей кластеров $v[1], v[2], \dots v[k]$.
5. Для $\forall$ v' определим кластер i, $i \in 1 \dots k$: $I_{r,r'[E]}(v[i] \cdot v') \to \max\limits_{v_i}$, получим σ'.
6. Для $\forall$ вершины v определим наиболее вероятное сообщество σ''_v, исходя из полученной σ'.
7. Получим новые оценки p и Q.
8. Для $\forall$ вершины v определим наиболее вероятное сообщество, исходя из σ''.
9. **Выход:** метка каждой вершины v.

Пример SSBM $n = 1000$, $k = 5$

$n = 1000$, $k = 5$, $p = (0.2, 0.2, 0.2, 0.2, 0.2)$,

$$W = \begin{pmatrix} 0.02 & 0.001 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.02 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.02 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.02 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.001 & 0.02 \end{pmatrix}$$

.

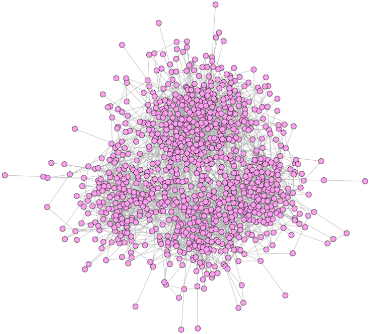

Рис.: Граф, $n = 1000$.

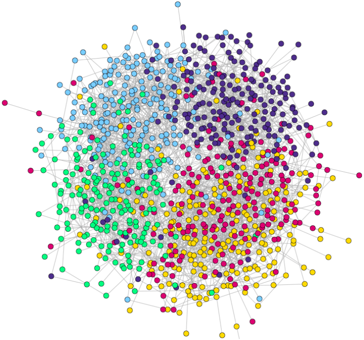

Рис.: Размеченный граф, $k = 5$.

Пример SSBM $n = 12,\ k = 2$

Параметрами алгоритма является вектор $p = (0.5, 0.5)$, матрица W — симметричная и имеет следующую структуру:

$$W = \begin{pmatrix} 0.66 & 0.16 \\ 0.16 & 0.66 \end{pmatrix}$$

.

Проверяем условие разделимости симметричной модели

$$W_{ij} = \begin{cases} a/n, & \text{если } i = j, \\ b/n, & \text{иначе.} \end{cases}$$

Следовательно $a = 0.66 \cdot n = 7.92$, $b = 0.16 \cdot n = 1.92$.

Тогда SNR $= \frac{(a-b)^2}{k(a+(k-1)b)} = 1.82 > 1$.

Пример SSBM $n = 12,\ k = 2$

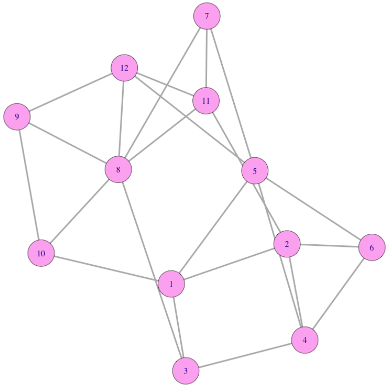

Рис.: Кластеризуемый граф, $n = 12$.

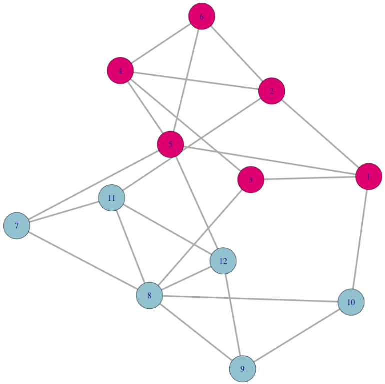

Рис.: Кластеризуемый граф, $n = 12,\ k = 2$.

Результаты

- В работе дан обзор литературы, освещающей актуальность и проблематику анализа больших данных.
- Представлены характеристики больших данных и некоторые методы их кластеризации.
- Приведено подробное конструктивное описание алгоритма SBM.
- Реализован алгоритм стохастической блочной модели и симметричной стохастической блочной модели.
- Программно реализована проверка разделимсти графа.
- Представлена визуализация полученной кластерной структуры графа.