Двухэтапный метод для выявления кластерной структуры сети (графа связей)

Авторы: Миркин Б.Г., Неваленная Ю.В.

ФКН, НИУ ВШЭ

Москва, 2019

Критерии кластеризации

- $f(S) = \sum_{i,j \in S} a_{ij}$ - суммарный критерий внутрикластерного сходства объектов (1);
- $f(S, \pi) = \sum_{i,j \in S} (a_{ij} - \pi)$ - модифицированный критерий (1) (Куперштох, Трофимов, Миркин 1976) (2);

где S – кластер на рассматриваемом множестве объектов I, a_{ij} – элемент матрицы сходства A, π - заданный константный порог.

Критерии кластеризации

Свойство 1

Если S является локально-оптимальным кластером по критерию (2), то он обладает следующим свойством:

$$\begin{cases} a(k,S) > \pi, k \in S \\ a(k,S) < \pi, k \notin S \end{cases}$$

где $a(k,\ S) = \frac{\sum_{i \in S} a_{ik}}{|S|-1}$ – величина, отражающая среднее сходство объекта k с объектами из S ($|S|$ - мощность кластера).

Критерий полусредней связи

- $a(S) = \frac{\sum_{i,j \in S} a_{ij}}{|S|(|S|-1)}$ - критерий средней связи между объектами кластера S (3);
- $g(S) = \frac{\sum_{i,j \in S} a_{ij}}{|S|}$ - критерий полусредней связи (4);
- $\Delta(S,k) = g(S \pm k) - g(S) = z_k \frac{(|S|+z_k)[a(S)-2a(k,S)]}{(|S|+1)}$ - изменение критерия $g(S)$ в зависимости от добавления/удаления объекта k (5);

где z - вектор принадлежности объектов кластеру: $z_i = 1$, если $i \in S$ и $z_i = -1$ – иначе.

Критерий полусредней связи

Свойство 2

Если S является локально-оптимальным кластером по критерию (4), то он обладает следующим свойством:

$$\begin{cases} a(k,S) > \frac{a(S)}{2}, k \in S \\ a(k,S) < \frac{a(S)}{2}, k \notin S \end{cases}.$$

где $a(k,\ S) = \frac{\sum_{i \in S} a_{ik}}{|S|-1}$ – величина, отражающая среднее сходство объекта k с объектами из S.

Алгоритм ВК

На вход: $A + A'$ (предобработанная матрица); начальный объект $i \in I$. **На выходе:** кластер S; его вклад в разброс данных $a^2(S)|S|(|S|-1)/(A,A)$ и интенсивность $a(S)$.

1. *Инициализация:* $\vec{z}$, такой что
$$z_j = \begin{cases} 1, & \text{if } j = i \text{ or } j = argmax(a_{ij}) \\ -1, & \text{otherwise} \end{cases} \quad \forall j \in I$$
Если $max(a_{ij}) < 0$ - переход к п. 4. Рассчитывается значение критерия $g(S)$ (4).
2. *Основной шаг:* Для $\forall k \in I$ рассчитывается значение $\Delta(S,k)$ (5); выбирается $\Delta(S,k^*) = \max(\Delta(S,k))$.
3. *Проверка:* если $\Delta(S,k^*) > 0$, то в кластер добавляется (или из кластера удаляется) объект k^*: $z_{k^*} = -z_{k^*}$. Пересчитывается $g(S)$. Если же $\Delta(S,k^*) \leq 0$ или $|S| < 2$ - переход к п. 4. Иначе, возврат к п. 2.
4. *Вывод результата.*

Алгоритм ВКС

1. *Сбор данных:* в цикле по объектам $i \in I$ запускается алгоритм ВК.
2. *Инициализация:* создается матрица сходства между полученными кластерами C, такая что: $c_{ij} = \frac{|i|+|j|}{2|i||j|} |i \cap j|$.
3. *Основной шаг:* повторяется п. 1, но уже для множества кластеров с использованием матрицы C.
4. *Проверка:* на данном этапе получаются кластеры, содержащие кластеры (полученные п. 1), пересекающиеся большими своими частями. Из каждого кластера выбирается его представитель, а именно тот из первоначальных кластеров, который имеет максимальную величину вклада. Если требуется дополнительноая фильтрация, выполняется переход к п. 2, иначе переход к п. 5.
5. *Вывод результата.*

Предварительное преобразование данных

- Вычитание среднего значения:
  $avg(A) = \frac{\sum_{1 \le i \le N, 1 \le j \le N} a_{ij}}{N(N-1)}$, число элементов матрицы = N^2;
- *Модулярный метод*, вычитание случайного взаимодействия: $p_{ij} = \frac{a_{i+} a_{+j}}{a_{++}}$, $a_{i+} = \sum_{k \in I} a_{ik}$, $a_{+j} = \sum_{k \in I} a_{kj}$, $a_{++} = \sum_{i,j \in I} a_{ij}$.

Данные Евровидения

- 1975-1997 гг. (20 век), всего 35 стран, период “голоса жюри”
- 1998-2015 гг. (21век), отобраны страны, выступавшие более 9 раз за период, всего 35 стран, период “голоса народа”
- Все время: 1975-2015 гг., отобраны страны, выступавшие более 5 раз за период, всего 48 стран

12-балльная система оценивания и учитываются голоса только в финалах конкурса.
Сформированы несимметричные матрицы сходства $A = (a_{ij})$, где a_{ij} - среднее число баллов, выставленное представителями страны i участникам страны j за данный промежуток времени.

Результаты (1975-2015)

Рис.1 Спектральный метод + вычитание среднего

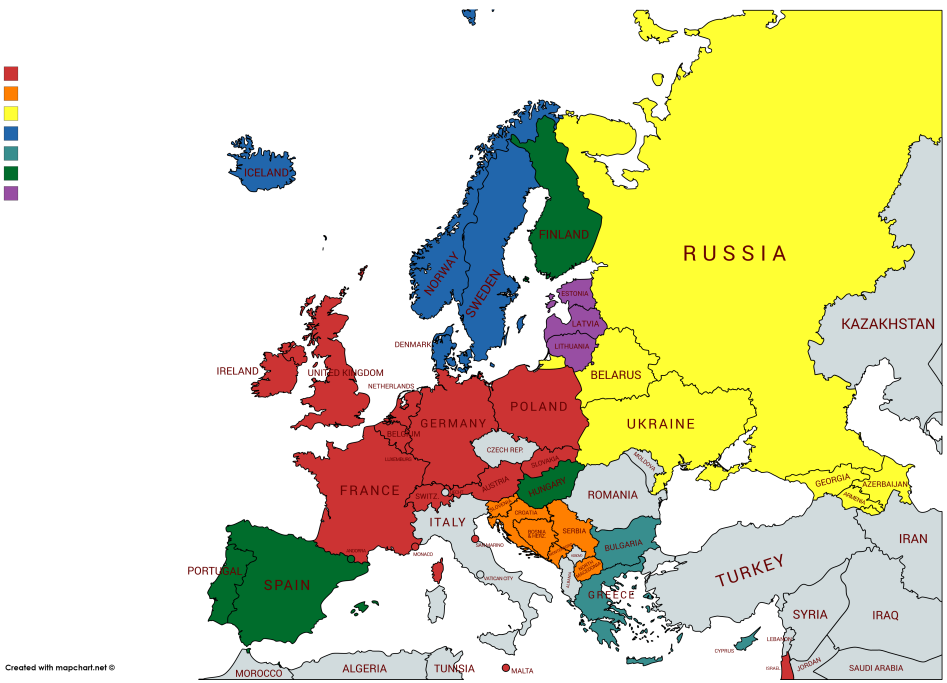
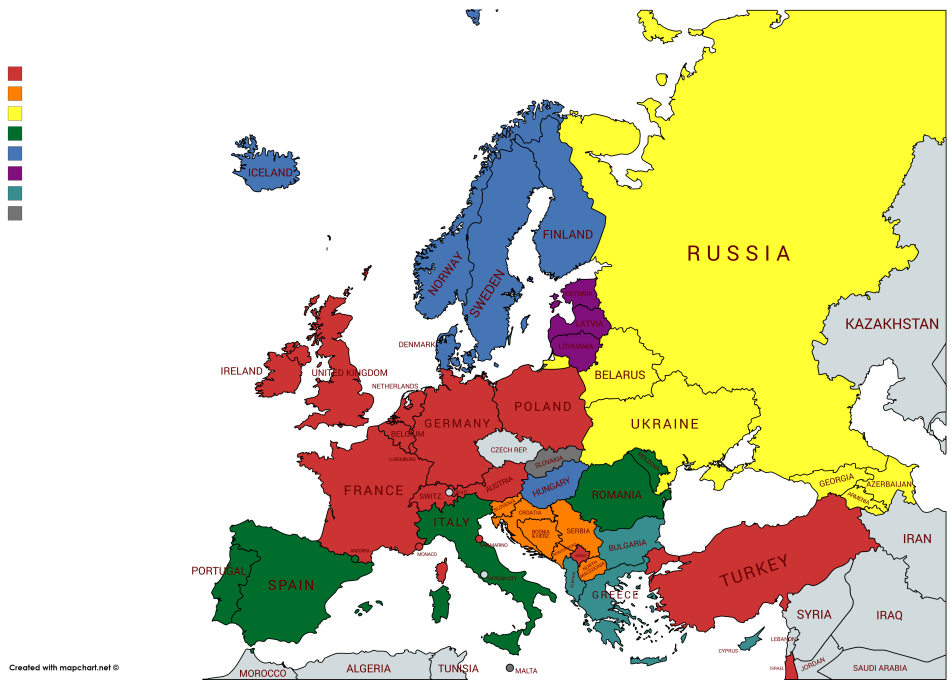


Рис.2 Спектральный + модулярный метод



Результаты (1975-2015)

Рис.3 ВКС + вычитание среднего

Рис.4 ВКС + модулярный метод

Результаты (1975-2015)

Рис.5 ВКС + модулярный метод (промежуточный этап)

Рис.6 (Рис. 4) ВКС + модулярный метод

Результаты

- **1975-1997:** 1) Южная, 2) Северная и Восточная и 3) Западная Европа (слабая предвзятость)
- **1998-2015:** 1) Северный, 2) Балканский (+ Турция) и 3) Постсоветский блоки (явная "положительная" предвзятость телезрителей, не по всей Европе, а на определенных участках)
- **1975-2015:** 1) Балканский блок, 2) Постсоветский блок, 3) Кипр, Греция, Болгария, 4) Северная и Западная Европа (наиболее нечеткий блок)

Спасибо за внимание!