Двухкомпонентная функция качества кластеризации множества элементов, представленных парными сравнениями

Двоенко С.Д.

Тульский государственный университет

ИОИ-2014

30 слайдов

План презентации

- Беспризнаковый алгоритм *k*-means
  - по расстояниям
  - по близостям
- Перестановочный *k*-means (без средних)
- Двухкомпонентная целевая функция для беспризнакового перестановочного *k*-means (bi-meanless *k*-means)
  - прямая форма
  - двойственная форма
  - снижение вычислительной сложности
- Эксперименты и результаты

Несмещенное разбиение

- Кластер-анализ:

  $\omega_i \in \Omega, i = 1, \ldots N, \mathbf{x}_i = (x_{i1}, \ldots x_{in}), X(N, n)$
- Алгоритм *k-means* основан на идее несмещенного разбиения:

  Каждый кластер $\Omega_k, k = 1, \ldots K$ представлен его "представителем" $\tilde{\mathbf{x}}_k$ и центр каждого кластера представлен его "средним" объектом $\overline{\mathbf{x}}_k$

Несмещенное разбиение

- Если представители и центры совпадают для всех кластеров $\tilde{\mathbf{x}}_k = \overline{\mathbf{x}}_k$, то несмещенное разбиение найдено
- Если нет, то разбиение - смещенное
- Тогда следует назначить средние объекты представителями и заново переклассифицировать все объекты на основе их минимальных расстояний до представителей

Отсутствие признаков

- Дана только матрица $D(N,N)$ расстояний
- Объект $\omega(\bar{\mathbf{x}}_k)$ как центр кластера в ней не представлен
- Можно считать объект, наиболее близкий к остальным в кластере, его центром $\bar{\omega}_k$
- Тогда при $\tilde{\omega}_k = \bar{\omega}_k$ будет получено несмещенное разбиение
- *Проблема*: в признаковом пространстве в общем случае будет получено смещенное разбиение, т.к. $\mathbf{x}(\bar{\omega}_k) \neq \bar{\mathbf{x}}_k$.

Алгоритм *k*-means

- *Теорема*: центр каждого кластера как его среднее арифметическое минимизирует его дисперсию и дисперсию всего разбиения:

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\mathbf{x}_i, \bar{\mathbf{x}}_k),$$

$$J(K) = \frac{1}{N}\sum_{k=1}^{K} N_k \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N}\sigma_k^2$$

Алгоритм *k*-means

- Дисперсия кластеров минимизируется для несмещенного разбиения и при отсутствии явно заданного пространства признаков:

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k),$$

$$J(K) = \frac{1}{N} \sum_{k=1}^{K} N_k \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N} \sigma_k^2$$

Алгоритм *k*-means

- Если множество Ω помещено в некоторое подходящее пространство признаков и $\mathbf{x}(\bar{\omega}_k) = \bar{\mathbf{x}}_k$, тогда два критерия

$$J^X(K) = \min_{\bar{\mathbf{x}}_1,\ldots\bar{\mathbf{x}}_K} J(K) \quad \text{и} \quad J^D(K) = \min_{\bar{\omega}_1,\ldots\bar{\omega}_K} J(K)$$

совпадают

$$J^X(K) = J^D(K)$$

- Но в общем случае

$$J^D(K) \geq J^X(K)$$

Алгоритм *k*-means

- Построим несмещенную кластеризацию, обеспечив выполнение условия

$$J^X(K) = J^D(K)$$

- Пусть $\omega_l \in \Omega$ расположен в начале координат (н.к.)
- $c_{ij} = (d_{li}^2 + d_{lj}^2 - d_{ij}^2)/2$ - скалярное произведение пары ω_i, ω_j, представленной расстоянием $d_{pq} = d(\omega_p, \omega_q)$, где $c_{ii} = d_{li}^2, \quad i = j$
- Главная диагональ матрицы $C_l(N, N)$ представляет квадраты расстояний до н.к. $\omega_l \in \Omega$
- Удобно поместить н.к. в центр «тяжести» множества $\omega_i \in \Omega, \ i = 1, \ldots N$ (*Torgerson, 1958*)

Алгоритм *k*-means

- Если н.к. совпадает с центром кластера, то этот центр $\bar{\omega}_k$ немедленно может быть представлен его расстояниями до всех остальных объектов

$$d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2;\ \ \omega_p, \omega_q \in \Omega_k,$$

$$\omega_i \in \Omega,\ \ i = 1, \ldots N$$

- Дисперсия кластера

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) =$$

$$\frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2.$$

Беспризнаковый *k*-means

- Беспризнаковый *k-means*:

Шаг 0. Назначить K начальных центров $\bar{\omega}_k^0$ и объявить их представителями $\tilde{\omega}_k^0,\ k=1,\ldots K.$

Шаг s. Перераспределить все объекты между кластерами:

1. Поместить ω_i в кластер $\omega_i \in \Omega_k^s$, если для $\omega_i \in \Omega_j^s$: $d(\omega_i, \bar{\omega}_k^s) \le d(\omega_i, \bar{\omega}_j^s),\ j=1,\ldots K,\ j \ne k.$
2. Пересчитать, если нужно, центры $\bar{\omega}_k^s, k=1,\ldots K$ и представить расстояниями $d(\omega_i, \bar{\omega}_k^s),\ i=1,\ldots N.$
3. Разместить следующий $i=i+1$ объект ω_i.
4. Стоп, если $\tilde{\omega}_k^s = \bar{\omega}_k^s, k=1,\ldots K$, иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s,$ $s=s+1.$

Перестановочный *k*-means

- Среднее квадратов расстояний между объектами в кластере вместе с расстояниями до самих себя

$$\eta'_k = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(\mathbf{x}_i - \mathbf{x}_j)^2 = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\mathbf{x}_i, \mathbf{x}_j)$$

- Очевидно, что $\eta'_k = 2\sigma_k^2 = 2\frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2$
- Обозначим $\eta_k = \eta'_k / 2 = \sigma_k^2$, где

$$\eta_k = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d^2(\omega_i, \omega_j)$$

Перестановочный *k*-means

- Дисперсия кластеризации также минимизируется

$$\tilde{J}(K) = J(K): \quad \tilde{J}(K) = \frac{1}{N}\sum_{k=1}^{K} N_k \eta_k = \sum_{k=1}^{K} \frac{N_k}{N} \eta_k$$

- Если объекты из Ω помещены в некоторое подходящее пространство и $\mathbf{x}(\bar{\omega}_k) = \bar{\mathbf{x}}_k$, если были вычислены, то критерии

$$\tilde{J}^X(K) = \min_{\Omega_1,\dots\Omega_K \in X} \tilde{J}(K) \text{ и } \tilde{J}^D(K) = \min_{\Omega_1,\dots\Omega_K \in D} \tilde{J}(K)$$

  также совпадут $\tilde{J}^X(K) = \tilde{J}^D(K)$

- Как и ранее, в общем случае $\tilde{J}^D(K) \geq \tilde{J}^X(K)$

Перестановочный *k*-means

- Перестановочный *k-means*:

Шаг 0. Определить K подмножеств $\Omega_k^0,\ k=1,...K$ и назначить их начальными.

Шаг s. Перераспределить все объекты между кластерами:

1. Поместить ω_i в кластер $\omega_i \in \Omega_k^s$ и $\tilde{J}^s(K)=\tilde{J}_k^s(K)$, если для $\omega_i \in \Omega_j^s$: $\tilde{J}_k^s(K)<\tilde{J}_j^s(K),\ j=1,...K,\ j\neq k.$
2. Разместить следующий $i=i+1$ объект ω_i.
3. Стоп, если объекты не перемещались, иначе $s=s+1.$

Разбиение по близостям

- Положительно полуопределенная матрица близостей $S(N,N)$ с элементами $s_{ij} = s(\omega_i, \omega_j) \geq 0$ может быть использована как матрица скалярных произведений в некотором пространстве размерности не более N
- Относительно некоторой точки $\omega_k \in \Omega$ как н.к. близости определяются как

$$s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2, \;\; s_{ii} = d_{ki}^2,$$

и расстояния определяются как

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$$

Разбиение по близостям

- Центр кластера $\bar{\omega}_k$ представлен его близостями с другими объектами

$$s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k} \sum_{p=1}^{N_k} s_{ip}, \quad \omega_p \in \Omega_k, \quad \omega_i \in \Omega, \quad i = 1, \ldots N$$

- Компактность кластера

$$\delta_k = \frac{1}{N_k} \sum_{i=1}^{N_k} s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k^2} \sum_{i=1}^{N_k} \sum_{p=1}^{N_k} s_{ip}, \quad \omega_i, \omega_p \in \Omega_k$$

Разбиение по близостям

- Несмещенное разбиение минимизирует дисперсию кластера σ_k^2 и максимизирует его компактность δ_k:

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} d_{ij}^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(s_{ii} + s_{jj} - 2s_{ij}) =$$

$$\frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} s_{ij} = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k$$

Разбиение по близостям

- Несмещенное разбиение минимизирует взвешенную среднюю дисперсию и максимизирует взвешенную среднюю компактность

$$J(K) = \sum_{k=1}^{K} \frac{N_k}{N} \sigma_k^2 = \sum_{k=1}^{K} \frac{N_k}{N} \left(\frac{1}{N_k} \sum_{i=1}^{N_k} s_{ii} - \delta_k \right) =$$

$$\frac{1}{N} \sum_{i=1}^{N} s_{ii} - \sum_{k=1}^{K} \frac{N_k}{N} \delta_k = c - \sum_{k=1}^{K} \frac{N_k}{N} \delta_k$$

- Взвешенная средняя компактность

$$I(K) = \sum_{k=1}^{K} \frac{N_k}{N} \delta_k, \quad I(K) = c - J(K)$$

k-means по близостям

- Беспризнаковый *k-means* по близостям :

Шаг 0. Назначить K начальных центров $\bar{\omega}_k^0$ и объявить их представителями $\tilde{\omega}_k^0,\ k=1,\ldots K.$

Шаг s. Перераспределить все объекты между кластерами:

1. Поместить ω_i в кластер $\omega_i \in \Omega_k^s$, если для $\omega_i \in \Omega_j^s$: $s(\omega_i, \bar{\omega}_k^s) \geq s(\omega_i, \bar{\omega}_j^s),\ j=1,\ldots K,\ j \neq k.$
2. Пересчитать, если нужно, центры $\bar{\omega}_k^s, k=1,\ldots K$ и представить их близостями $s(\omega_i, \bar{\omega}_k^s),\ i=1,\ldots N.$
3. Разместить следующий $i=i+1$объект ω_i .
4. Стоп, если $\tilde{\omega}_k^s = \bar{\omega}_k^s, k=1,\ldots K,$ иначе $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s,$ $s=s+1.$

Перестановочный *k*-means по близостям

- Перестановочный *k-means* по близостям:

Шаг 0. Определить K подмножеств $\Omega_k^0,\ k=1,...K$ и назначить их начальными.

Шаг s. Перераспределить все объекты между кластерами:

1. Поместить ω_i в кластер $\omega_i \in \Omega_k^s$ и $I^s(K) = I_k^s(K)$, если для $\omega_i \in \Omega_j^s$: $I_k^s(K) > I_j^s(K),\ j=1,...K,\ j \neq k.$
2. Разместить следующий $i = i+1$ объект ω_i .
3. Стоп, если объекты не перемещались, иначе $s = s+1.$

Двухкомпонентная ц.ф. для перестановочного *k*-means

- ***Прямая форма***: минимизировать внутри-кластерные дисперсии $\tilde{J}(K)$ и межкластерную близость $\delta(K)$: $\tilde{J}_\delta(K) = \tilde{J}(K) + \alpha\delta(K)$,

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} s_{pq}\ , \omega_p \in \Omega_k,\ \omega_q \in \Omega_l\ ,$$

  α - масштабирующий коэффициент

- 2ц.ф. *k-means* аналогичен беспризнаковому *k-means* по расстояниям для минимизации $\tilde{J}(K)$

Двухкомпонентная ц.ф. для перестановочного *k*-means

- ***Двойственная форма***: максимизировать внутрикластерный близости $I(K)$ и межкластерную дисперсию $\sigma^2(K)$: $I_\sigma(K) = I(K) + \alpha\sigma^2(K)$,

$$\sigma^2(K) = \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d_{pq}^2,\ \omega_p \in \Omega_k,\ \omega_q \in \Omega_l,$$

α - масштабирующий коэффициент

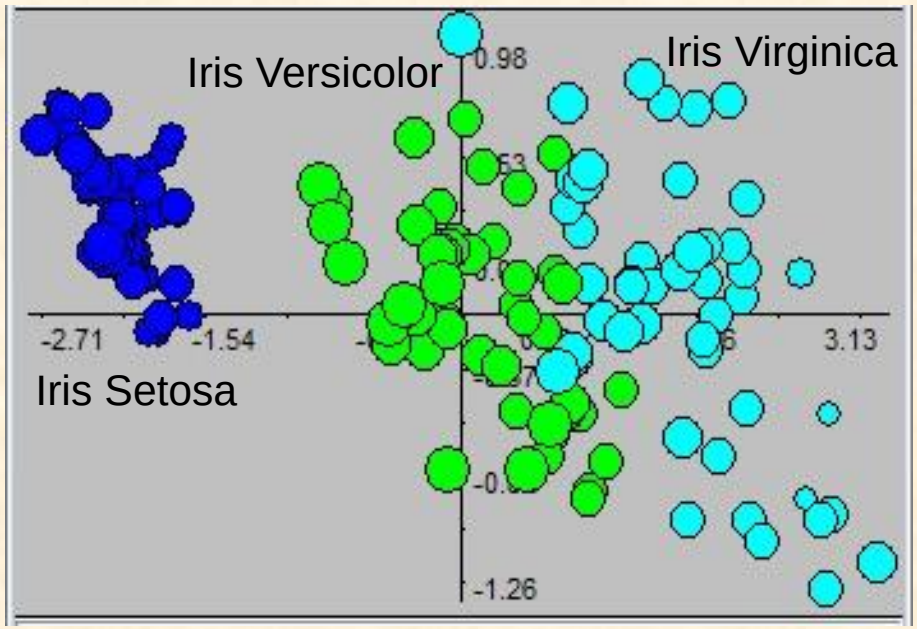
- 2ц.ф. *k-means* аналогичен беспризнаковому *k-means* по близостям для максимизации $I(K)$

Двухкомпонентная ц.ф. для перестановочного *k*-means

- Очевидно, что необходимо снизить вычислительную сложность рассмотренных алгоритмов на основе экономных формул для приращений целевой функции.
- Такие формулы существуют

Эксперименты. Ирисы

- Fisher R.A. “The Use of Multiple Measurements in Taxonomic Problems”, *Ann.Eugenics*, 7(9):179-188,1936.



Виды:
1. ***Iris Setosa*** (50 шт)
2. ***Iris Versicolor*** (50 шт)
3. ***Iris Virginica*** (50 шт)

Признаки:
1. Чашелистик длина
2. Чашелистик ширина
3. Лепесток длина
4. Лепесток ширина

Классы 2 и 3 слегка пересекаются

Данные показаны в пространстве первых трех г.к. (повернуто). Объясняемая доля дисперсии 99.5%

Эксперименты

- Подбор масштаба α для прямой формы ц.ф. $\tilde{J}_\delta(K) = \tilde{J}(K) + \alpha\delta(K)$ позволяет получить:

  1. улучшение результата для неразделимых кластеров: Versicolor (2) vs Virginica (3)

  2. улучшение результата для разделимых, но разных по размеру кластеров: небольшой Setosa (1) vs большой Versicolor и Virginica (2+3)

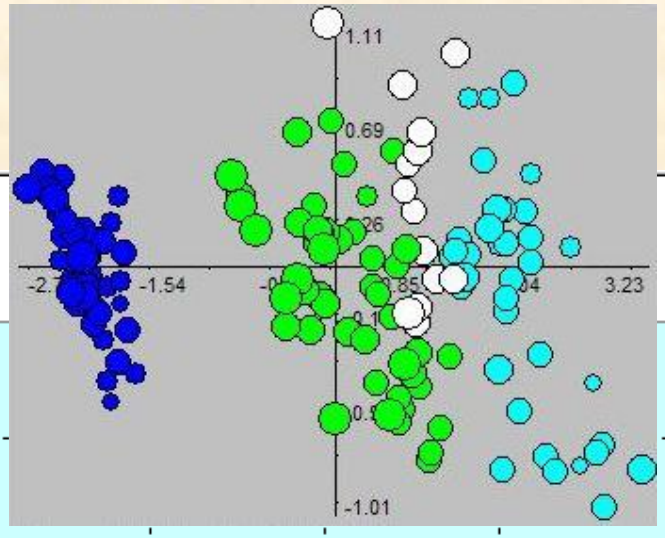
- Начальные разбиения были определены заранее, чтобы локальные свойства алгоритма *k*-means не влияли на результат

1. Неразделимые кластеры

3 кластера (Setosa-Versicolor-Virginica)				2 кластера (Versicolor-Virginica)			
Initial partition	Errors ($\alpha = 0$)	Optimal α	Errors (α_{opt})	Initial partition	Errors ($\alpha = 0$)	Optimal α	Errors (α_{opt})
50-50-50	16	3 – 6	15	50-50	16	12 –17.7	15
50-70-30	16	3 – 6	15	70-30	16	12 –17.7	15
50-30-70	16	3 – 6	15	30-70	16	12 – 17.7 22 – 22.4	15

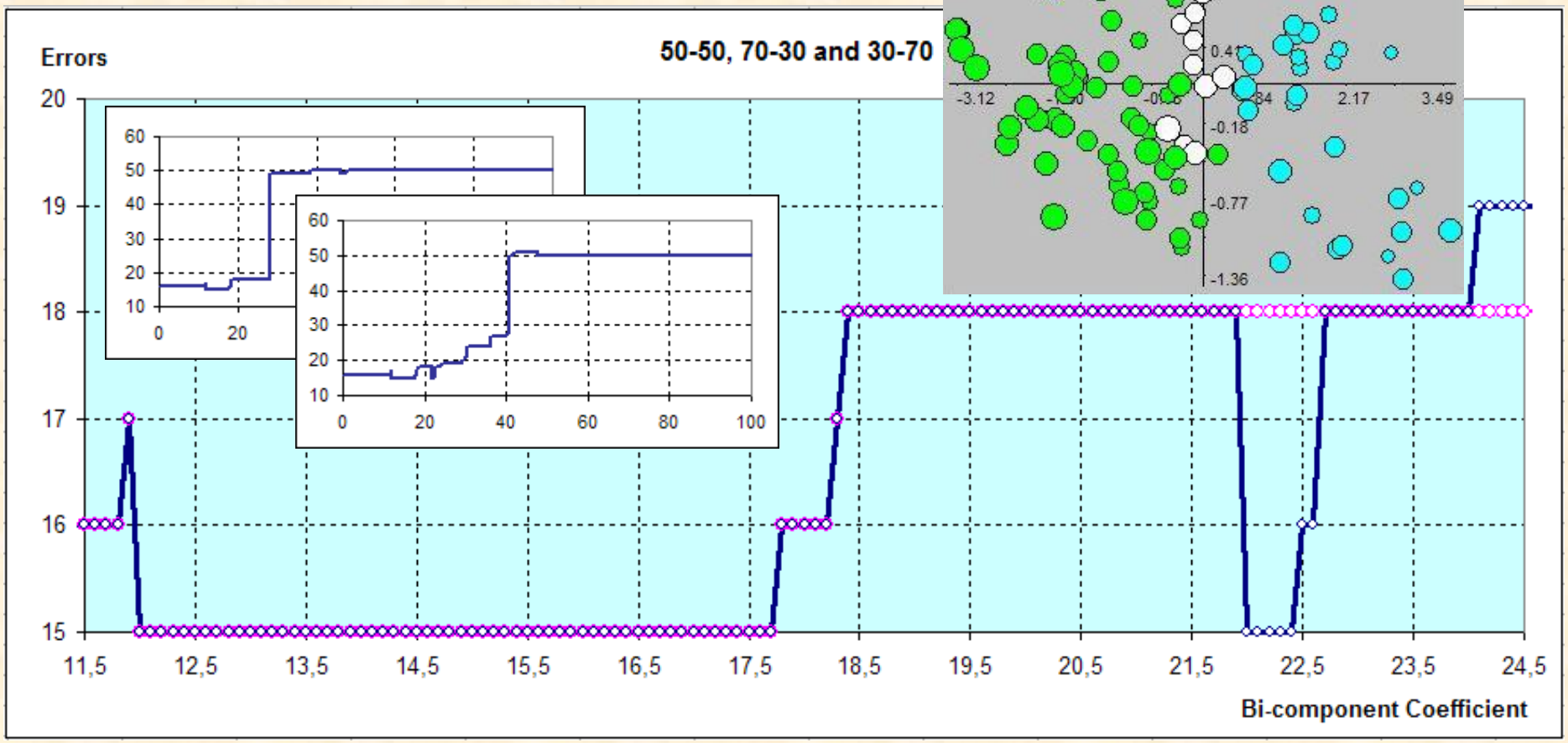
1. Неразделимые кластеры

Error Objects: 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 135, 139, 143, 147, 150



1. Неразделимые кластеры

Error Objects: 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 135, 139, 143, 147, 150

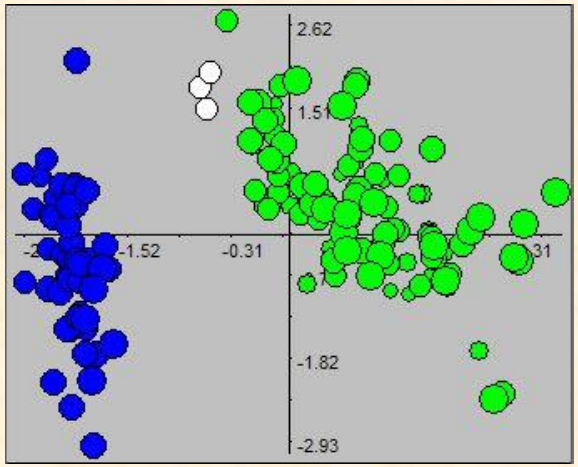


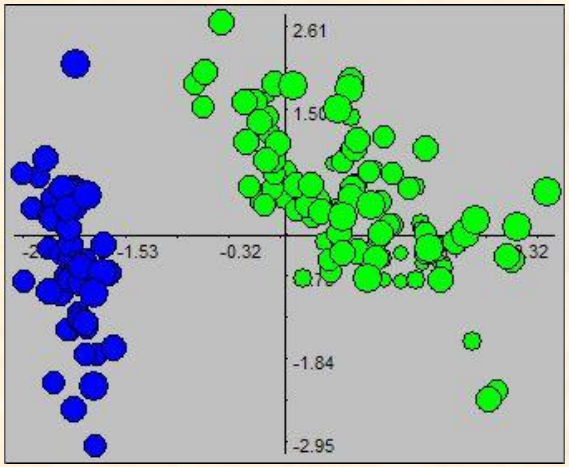
2. Разделимые кластеры

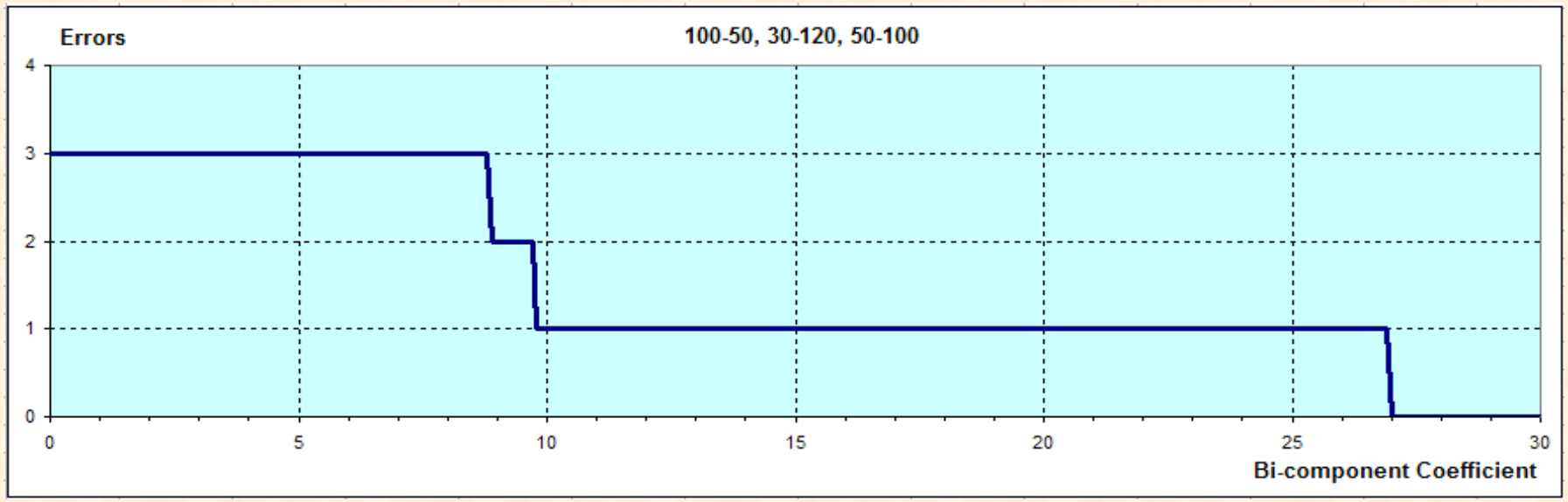
$\alpha = 0$ Errors = 3

$\alpha = 27$ Errors = 0

Error Objects: 58, 94,99







Заключение

- Перестановочный беспризнаковый алгоритм *k-means* обрабатывает квадратные матрицы парных сравнений так же, как и беспризнаковый алгоритм *k-means*
- Это позволяет при отсутствии признакового пространства даже не вводить понятие среднего объекта, т.к. его часто невозможно использовать в явном виде
- Предложенные процедуры гарантируют одинаковые результаты как при наличии пространства признаков, так и без него
- Алгоритм *k-means* с 2ц.ф. позволяет улучшить локальные свойства алгоритма *k-means*