Многомерный анализ данных. Спецкурс.

Лекция 2. Анализ нечисловых данных

ДВФУ/БСИ ДВО РАН

Кислов Д.Е.

28 ноября 2018 г.

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- непременимы многие статистические понятия (среднее, дисперсия и т.п.)

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- непременимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- непременимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- непременимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков
- попробовать ввести количественные меры «близости» объектов (учитывая специфику задачи)

Обработка нечисловых данных. Специфика задач.

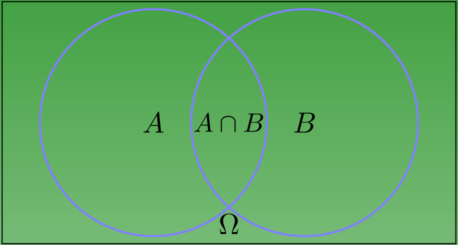
Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- непременимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков
- попробовать ввести количественные меры «близости» объектов (учитывая специфику задачи)
- оценивать вероятности наличия тех или иных признаков на основе предельных теорем

Меры сходства и различия

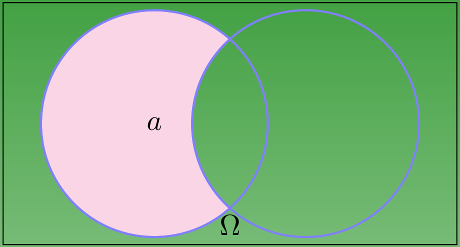


Основные меры:

	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$
\begin{aligned}
a &= |A - B| \\
b &= |B - A| \\
c &= |A \cap B| \\
d &= |\Omega - A \cup B|
\end{aligned}
$$

Меры сходства и различия

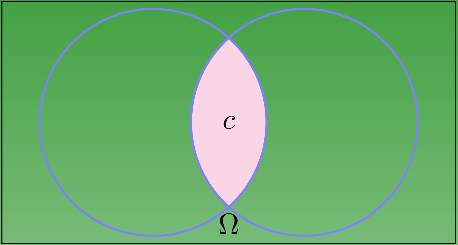


	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

Меры сходства и различия

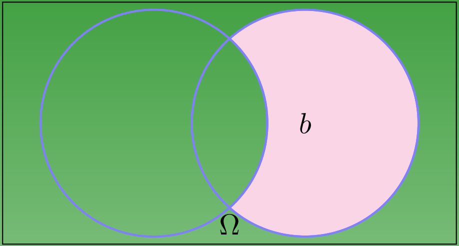


	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$
\begin{aligned}
a &= |A - B| \\
b &= |B - A| \\
c &= |A \cap B| \\
d &= |\Omega - A \cup B|
\end{aligned}
$$

Основные меры:

Меры сходства и различия

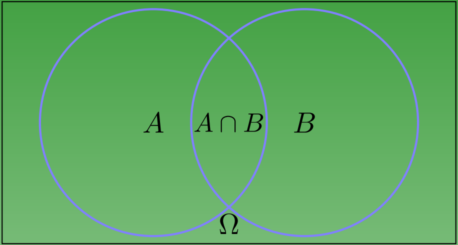


	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

Меры сходства и различия

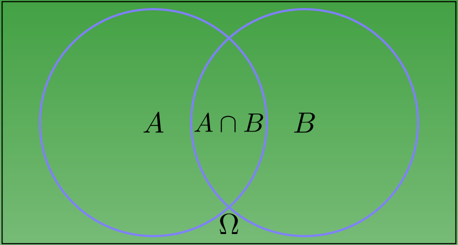


Основные меры:

	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Меры сходства и различия



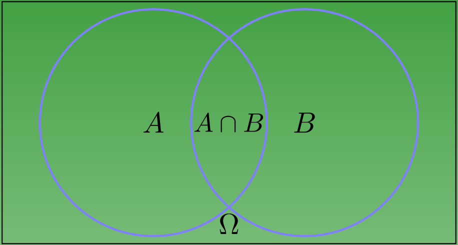
	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

- $\dfrac{c}{a+b+c}$ (Jaccard, 1901)

Меры сходства и различия

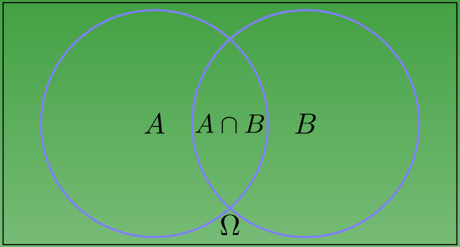


	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

- $\dfrac{c}{a+b+c}$ (Jaccard, 1901)
- $\dfrac{2c}{a+b+2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)



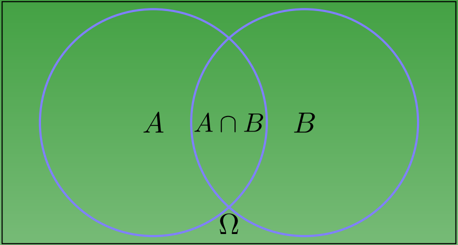
	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

- $\dfrac{c}{a+b+c}$ (Jaccard, 1901)
- $\dfrac{2c}{a+b+2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)
- $\dfrac{c}{a+b}$ (Кульчинский, 1927)

Меры сходства и различия



	B	$\overline{B}$	$\sum$
A	c	a	$a+c$
$\overline{A}$	b	d	$b+d$
$\sum$	$b+c$	$a+d$	$a+b+c+d$

$$a = |A - B|$$
$$b = |B - A|$$
$$c = |A \cap B|$$
$$d = |\Omega - A \cup B|$$

Основные меры:

- $\dfrac{c}{a+b+c}$ (Jaccard, 1901)
- $\dfrac{2c}{a+b+2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)
- $\dfrac{c}{a+b}$ (Кульчинский, 1927)
- $\dfrac{c}{c+a}, \dfrac{c}{c+b}$ (Шимкевич, 1926; Simpson, 1943)

Примеры вычислений

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$. Если A – стул, B – тренога, то $a = 1$ и $b = 0$.

Примеры вычислений

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$. Если A – стул, B – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \dfrac{3}{1+0+3} = 0.75$

Примеры вычислений

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если А – стул, В – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \dfrac{3}{1 + 0 + 3} = 0.75$
- мера Дайса $D = \dfrac{2 \cdot 3}{1 + 0 + 2 \cdot 3} = 6/7 \approx 0.857$

Примеры вычислений

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$. Если A – стул, B – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \dfrac{3}{1+0+3} = 0.75$
- мера Дайса $D = \dfrac{2 \cdot 3}{1+0+2 \cdot 3} = 6/7 \approx 0.857$
- мера Кульчинского $K = \dfrac{3}{1+0} = 3$

Примеры вычислений

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$. Если A – стул, B – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \dfrac{3}{1+0+3} = 0.75$
- мера Дайса $D = \dfrac{2 \cdot 3}{1+0+2\cdot 3} = 6/7 \approx 0.857$
- мера Кульчинского $K = \dfrac{3}{1+0} = 3$

Вычисления на R

```
library(sets)
set_similarity(set(1,2,3,4),set(1,2,5),method="Jaccard")
```

[1] 0.4

Параметризация мер сходства

Двухпараметрическое семейство мер (Б.И. Семкин, 2010):

$$K_{\tau;\eta} = \left(\frac{K_\tau^\eta(A,B) + K_\tau^\eta(B,A)}{2} \right)^{1/\eta},$$

$$K_\tau(A,B) = \frac{|A \cap B|}{(1+\tau)|A| - \tau|A \cap B|},$$

$$K_\tau(B,A) = \frac{|A \cap B|}{(1+\tau)|B| - \tau|A \cap B|}$$

$$-1 < \tau < \infty, -\infty < \eta < \infty$$

В этом случае $K_{0;-1}$ и $K_{1;-1}$ совпадают с коэффициентами Сёренсена-Дайса и Жаккара соответственно.

Параметризация мер сходства

Двухпараметрическое семейство мер (Б.И. Семкин, 2010):

$$K_{\tau;\eta} = \left(\frac{K_\tau^\eta(A,B) + K_\tau^\eta(B,A)}{2} \right)^{1/\eta},$$

$$K_\tau(A,B) = \frac{|A \cap B|}{(1+\tau)|A| - \tau|A \cap B|},$$

$$K_\tau(B,A) = \frac{|A \cap B|}{(1+\tau)|B| - \tau|A \cap B|}$$

$$-1 < \tau < \infty, -\infty < \eta < \infty$$

В этом случае $K_{0;-1}$ и $K_{1;-1}$ совпадают с коэффициентами Сёренсена-Дайса и Жаккара соответственно.

Вывод

Используемые меры имеют много общего, они в определенном смысле «эквивалентны»

Точный тест Фишера

Задача

Исследуется вопрос об эффективности обработки с целью последующего проращивания жёлудей. В результате эксперимента построена следующая таблица сопряженности:

	не взошел	взошел	$\sum$
обработано	1	10	11
не обработано	4	3	7
$\sum$	5	13	18

Целесообразно ли применение данной обработки, или «увеличение» всхожести в результате обработки вполне могло возникнуть случайно?

Точный тест Фишера

Задача

Исследуется вопрос об эффективности обработки с целью последующего проращивания жёлудей. В результате эксперимента построена следующая таблица сопряженности:

	не взошел	взошел	$\sum$
обработано	1	10	11
не обработано	4	3	7
$\sum$	5	13	18

Целесообразно ли применение данной обработки, или «увеличение» всхожести в результате обработки вполне могло возникнуть случайно?

Решение

Нужно вычислить вероятность реализации таблицы $[(1, 10), (4, 3)]$, а также более «худшего» варианта, $[(0, 11), (5, 2)]$, т.е. когда после обработки вообще все семена взошли. Если сумма этих вероятностей будет мала, то, вероятно, что обработка (а не случайность) определяет исход прорастания.

Решение

$$\frac{C_5^1 \cdot C_{13}^{10}}{C_{18}^{11}} + \frac{C_5^0 C_{13}^{11}}{C_{18}^{11}} =$$

$$\frac{1430}{31824} + \frac{78}{31824} \approx 0.047$$

Точный тест Фишера

Решение

$$\frac{C_5^1 \cdot C_{13}^{10}}{C_{18}^{11}} + \frac{C_5^0 C_{13}^{11}}{C_{18}^{11}} =$$

$$\frac{1430}{31824} + \frac{78}{31824} \approx 0.047$$

Таким образом, вероятность наблюдать исход эксперимента, или даже исход, когда все желуди взошли вследствие случая (а не действия обработки), равна около 4.7%; это весьма маленькое значение, поэтому результаты наблюдений следует интерпретировать, что имеет место значимое влияние обработки на результат прорастания желудей.

Точный тест Фишера

Общий случай

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

$$P(a,b;c,d) = \frac{C_{a+c}^{a} \cdot C_{b+d}^{b}}{C_{n}^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$

$$n = a+b+c+d$$

Точный тест Фишера

Общий случай

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

$$P(a,b;c,d) = \frac{C_{a+c}^{a} \cdot C_{b+d}^{b}}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$
$$n = a+b+c+d$$

Односторонний тест: «усугубление» наблюдаемой ситуации

$\sum_{j=0}^{a} P(j,\tilde{b};\tilde{c},\tilde{d})$, при условии: $j+\tilde{c} = a+c$, $\tilde{b}+\tilde{d} = b+d$,
$\tilde{b}+j = a+b$, $j+\tilde{b}+\tilde{c}+\tilde{d} = n$

Точный тест Фишера

Общий случай

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

$$P(a,b;c,d) = \frac{C_{a+c}^{a} \cdot C_{b+d}^{b}}{C_{n}^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$

$$n = a+b+c+d$$

Двусторонний тест

Точный тест Фишера

Общий случай

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

$$P(a,b;c,d) = \frac{C_{a+c}^{a} \cdot C_{b+d}^{b}}{C_{n}^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$
$$n = a+b+c+d$$

Двусторонний тест: что считать «усугублением»?

Точный тест Фишера

Общий случай

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

$$P(a,b;c,d) = \frac{C_{a+c}^{a} \cdot C_{b+d}^{b}}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$
$$n = a+b+c+d$$

Двусторонний тест: что считать «усугублением»?

$\sum\limits_{\tilde a,\tilde b,\tilde c,\tilde d} P(\tilde a,\tilde b;\tilde b,\tilde d)$, суммирование при условиях:

$P(\tilde a,\tilde b;\tilde c,\tilde d) \le P(a,b;c,d)$, $\tilde a+\tilde b = a+b$, $\tilde c+\tilde d = c+d \ldots$

Точный тест Фишера: приближенные вычисления

	не взошел	взошел	$\sum$
обработано	a	b	$a+b$
не обработано	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

Гипотеза: наблюдаемое распределение a, b, c, d результат случая

Аппроксимация распределением χ^2 (с поправкой Ейтса)

$$\chi^2_{\text{выч.}} = \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$n = a+b+c+d$$

условия применимости: $a, b, c, d \geq 5, n \geq 40$

Гипотеза отвергается на уровне значимости α, если $\chi^2_{\text{выч.}} > \chi^2_{1-\alpha}(1)$ (в частности, $\chi^2_{0.95}(1) \approx 3.85$, 1 – число степеней свободы для таблицы 2×2)

Задачи классификации

- Классификация в отсутствии обучающей выборки (кластеризация);

Задачи классификации

- Классификация в отсутствии обучающей выборки (кластеризация);
- Классификация при наличии обучающей выборки (по прецедентам);

- Иерархический (агломеративная, дивизивные);

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация
- Вероятностные

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация
- Вероятностные
- Нейросетевые

Общая структура алгоритмов

- задание расстояние между кластеризуемыми объектами;
- задание расстояние между группами объектов;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum\limits_j (x_j - y_j)^2$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

- Евклидово расстояние: $\rho(x,y) = \sum_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x,y) = \max_j |x_j - y_j|$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \dots, x_n$ и $y_1, \dots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x, y) = \max_j |x_j - y_j|$;
- Расстояние city-block: $\rho(x, y) = \sum_i |x_i - y_i|$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x, y) = \max_j |x_j - y_j|$;
- Расстояние city-block: $\rho(x, y) = \sum_i |x_i - y_i|$;
- Расстояние Минковского: $\rho(x, y)^p = \sum_i (x_i - y_i)^p$;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);

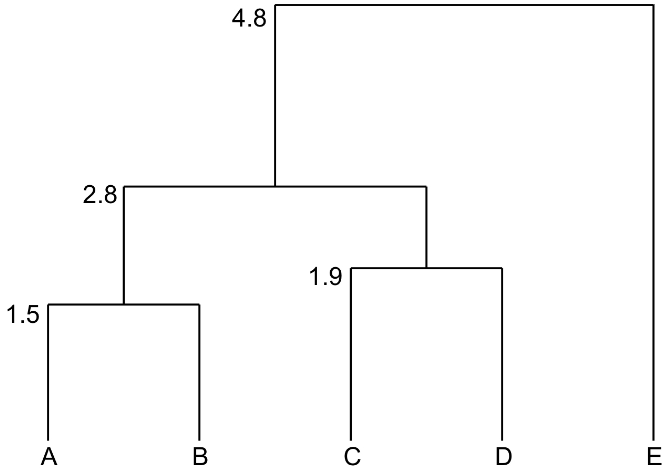
Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;
- центроидный метод;

Представление иерархической кластеризации в виде дендрограммы



Метод k-средних

- задается число кластеров k;
- в факторном пространстве случайным образом выбираются начальные приближения центров кластеров;
- ближайшие к j-му центру точки помещаются в j-й кластер;
- пересчитываются центроиды кластеров;

Последние 2 шага повторяются пока алгоритм не сойдется.

Сравнение кластерных структур

Задача

Можно ли построить какую-либо меру, чтобы сравнить, например, кластерные структуры?:

$$\begin{array}{c} a, a, a, b, b, c, c, c, c \\ 1, 1, 1, 3, 3, 2, 2, 2, 2 \end{array}$$

или

$$\begin{array}{c} a, a, a, b, b, c, c, c, c \\ 2, 2, 2, 1, 1, 3, 3, 3, 1 \end{array}$$

Индекс Рэнда

Пусть $X = (x_1, \ldots, x_n)$ – n-элементное множество;
$P = (A_1, \ldots, A_r)$ и $Q = (B_1, \ldots, B_s)$ – два его разбиения.
Определим

- a – число пар элементов попавших в один кластер в разбиениях P и Q одновременно;
- b – число пар элементов, находящихся в разных кластерах в разбиениях P и Q;
- c – число пар элементов, находящихся в одном кластере в P разбиении, но в разных в Q;
- d – число пар элементов, находящихся в разных кластерах в P разбиении, но в одном в Q;

В этом случае $a + b$ – характеризует степень совпадения кластеров, если $c = d = 0$, то кластерные структуры совпадают.
Индекс Рэнда:

$$I_R = \frac{a+b}{a+b+c+d} = \frac{a+b}{C_n^2}$$