

О МЕРЕ СВЯЗИ ЗНАЧЕНИЙ ПРИЗНАКОВ

Ю.Г. Дмитриев, Е.В. Самойлова

На основе оценки с использованием априорной информации, имеющей наименьшую дисперсию в классе линейных несмещенных оценок, вводится мера связи между отдельными значениями качественных признаков. Исследована связь введенной меры и стандартного коэффициента сопряженности. Показано, как меняется величина меры в зависимости от выбора значений.

При работе с качественными социально-экономическими данными обычно исследуются зависимости между признаками в целом, а не между их конкретными значениями. Однако описание социальных закономерностей базируется на описании связей между отдельными значениями признаков (конкретными свойствами объектов), а не признаками в целом [1]. Примером может служить исследование, цель которого - узнать, какие именно социальные, экономические или демографические характеристики потребителей сильнее всего влияют на то, приобретают ли они товары в тех или иных магазинах.

Кроме того, даже когда между какими-то отдельными значениями признаков имеется достаточно сильная связь, между признаками в целом связь может быть слабой [1], [3]. Поэтому представляется естественным ввести меру связи между отдельными значениями качественных признаков.

Построение меры связи на основе оценки, учитывающей априорную информацию

Рассмотрим конечное множество $\Omega = \{O_1, \dots, O_N\}$, состоящее из N объектов O_j , каждый из которых характеризуется S качественными признаками $C_1, \dots, C_S$. Признак C_s принимает $m_s > 1$ значений c_s^k ($k = \overline{1, m_s}, s = \overline{1, S}, \sum_{s=1}^S m_s = M$).

Пусть Q_s^k - множество объектов из Ω , для которых признак C_s принимает значение c_s^k , $\nu(Q_s^k)$ - число элементов этого множества. Тогда их доля во всей совокупности равна

$$P_N(Q_s^k) = \frac{\nu(Q_s^k)}{\nu(\Omega)} = \frac{\nu(Q_s^k)}{N}.$$

Выберем $m < M$ других значений признаков $C_1, \dots, C_S$ и для простоты переобозначим эти значения $b_1, \dots, b_m$, а соответствующие им множества объектов обозначим B_i ($i = \overline{1, m}$). Значения b_i могут принадлежать как одному

и тому же, так и разным признакам; потребуем лишь, чтобы $\nu\left(\bigcap_{i=1}^m B_i\right) < N$. Доля объектов множества B_i во множестве Ω

$$P_N(B_i) = \frac{\nu(B_i)}{N}.$$

Обозначим $a = c_s^k$, $A = Q_s^k$. Допустим, что доля $P_N(A)$ нам неизвестна, но известны $P_N(B_i)$ ($i = \overline{1, m}$).

Пусть имеется выборка без возвращения объема n объектов из Ω . Требуется по данной выборке оценить $P_N(A)$, используя информацию о значениях $P_N(B_i)$. Рассмотрим оценку следующего вида:

$$\begin{aligned} P_n^*(A) &= P_n(A) - \sum_{i=1}^m \lambda_i (P_n(B_i) - b_i) = \\ &= P_n(A) - \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{bmatrix}^T \begin{bmatrix} P_n(B_1) - P_N(B_1) \\ \vdots \\ P_n(B_m) - P_N(B_m) \end{bmatrix}, \end{aligned} \quad (1)$$

где $P_n(A) = \nu_n(A)/n$, $\lambda_1, \dots, \lambda_m$ - числовые коэффициенты.

Данная оценка является несмещенной, т.е. математическое ожидание $MP_n^*(A) = P_N(A)$. Коэффициенты $\lambda_1, \dots, \lambda_m$ выбраны из условия минимума дисперсии $DP_n^*(A)$ оценки и имеют вид

$$[\lambda_1 \ \dots \ \lambda_m]^T = K_{BB}^{-1} K_{AB}, \quad (2)$$

где

$$K_{AB} = \begin{bmatrix} P_N(AB_1) - P_N(A)P_N(B_1) \\ P_N(AB_2) - P_N(A)P_N(B_2) \\ \vdots \\ P_N(AB_m) - P_N(A)P_N(B_m) \end{bmatrix},$$

K_{BB} - симметричная матрица вида

$$\begin{bmatrix} P_N(B_1) - P_N^2(B_1) & \Lambda & P_N(B_1 B_m) - P_N(B_1)P_N(B_m) \\ P_N(B_1 B_2) - P_N(B_1)P_N(B_2) & \Lambda & P_N(B_2 B_m) - P_N(B_2)P_N(B_m) \\ \vdots & \vdots & \vdots \\ P_N(B_1 B_m) - P_N(B_1)P_N(B_m) & \Lambda & P_N(B_m) - P_N^2(B_m) \end{bmatrix}$$

(K_{BB} предполагается невырожденной),
 $P_N(AB_i)$ - доля объектов совокупности, одновременно принадлежащих A и B_i , $P_N(B_i B_j)$ - доля объектов совокупности, одновременно принадлежащих B_i и B_j .

При этом дисперсия

$$DP_n^*(A) = \frac{N-n}{N-1} \left(\frac{P_N(A) - P_N^2(A)}{n} - \frac{K_{AB}^T K_{BB}^{-1} K_{AB}}{n} \right) =$$

$$= DP_n(A) - \left(\frac{N-n}{N-1} \right) \frac{K_{AB}^T K_{BB}^{-1} K_{AB}}{n}, \quad (3)$$

где $DP_n(A) = \frac{(N-n)(P_N(A) - P_N^2(A))}{(N-1)n}$.

Поскольку квадратичная форма $K_{AB}^T K_{BB}^{-1} K_{AB}$ неотрицательно определена, то

$$DP_n^*(A) \leq DP_n(A). \quad (4)$$

Поделив равенство (3) на $DP_n(A)$, получим

$$\frac{DP_n^*(A)}{DP_n(A)} = 1 - \frac{K_{AB}^T K_{BB}^{-1} K_{AB}}{P_N(A) - P_N^2(A)}. \quad (5)$$

Обозначим

$$\eta_m(a) = \eta_m(a | b_1, K, b_m) = \frac{K_{AB}^T K_{BB}^{-1} K_{AB}}{P_N(A) - P_N^2(A)}. \quad (6)$$

Величина $\eta_m(a)$ характеризует степень влияния значений $b_1, \dots, b_m$ на точность оценивания (по величине дисперсии) $P_N(A)$. Поэтому можно рассматривать $\eta_m(a)$ как меру связи между значением a признака C_s и набором значений $b_1, \dots, b_m$ признаков $C_1, \dots, C_S$.

Свойства меры связи

Рассмотрим некоторые свойства введенной меры связи $\eta_m(a)$.

1) $0 \leq \eta_m(a) \leq 1$. Это следует из соотношений (4)-(6). Чем ближе $\eta_m(a)$ к единице, тем меньше величина дисперсии оценки $P_n^*(A)$. Поэтому будем говорить, что чем ближе $\eta_m(a)$ к единице, тем сильнее связь между значением a и значениями $b_1, \dots, b_m$.

Величина $\eta_m(a) = 0$, когда для всех $B_i, i = \overline{1, m}$ выполняется равенство $P_N(AB_i) = P_N(A)P_N(B_i)$.

Величина $\eta_m(a) = 1$, когда для всех значений $b_1, \dots, b_m$ выполняется:

- $\forall b_i, b_j : i \neq j \quad P_N(B_i B_j) = 0$;
- $P_N(AB_i) = P_N(B_i), \quad i = \overline{1, m}$

(либо $P_N(AB_i) = 0, \quad i = \overline{1, m}$);

- $\sum_{i=1}^m P_N(B_i) = P_N(A)$

(либо $\sum_{i=1}^m P_N(B_i) = 1 - P_N(A)$)

2) По аналогии тому, как это сделано в [2], можно показать, что

$$\eta_{m+1}(a | b_1, \dots, b_m, b_{m+1}) \geq \eta_m(a | b_1, \dots, b_m).$$

Однако если взять другой набор $b'_1, \dots, b'_m$, в котором хотя бы одно значение $b'_i \neq b_i (i = \overline{1, m})$, величина $\eta_{m+1}(a | b'_1, \dots, b'_m, b_{m+1})$ может быть как больше, так и меньше $\eta_m(a | b_1, \dots, b_m)$.

Таким образом, величина данной меры связи зависит как от количества m выбранных значений признаков $b_1, \dots, b_m$, так и от того, какие именно значения b_i выбраны.

3) В случае $m=1$ (выбрано только одно значение $b_1=b$):

$$\eta_1(a | b) = \eta_1(b | a) = \eta_1(\bar{a} | \bar{b}) = \eta_1(a | \bar{b}) = \eta_1(\bar{a} | \bar{b}) =$$

$$= \frac{(P_N(AB) - P_N(A)P_N(B))^2}{(P_N(A) - P_N^2(A))(P_N(B) - P_N^2(B))},$$

где $\bar{a}, \bar{b}$ - значения, которым соответствуют множества объектов $\bar{A} = \Omega \setminus A$ и $\bar{B} = \Omega \setminus B$.

Связь с коэффициентом Пирсона χ^2

В некоторых ситуациях, когда количество значений признаков $m_i > 2$, связь между признаками в целом может быть слабой, даже если между их отдельными значениями есть достаточно сильная связь. В [3] приведены примеры таких ситуаций.

В [1] был предложен метод измерения связи между конкретными значениями a_k и b_i двух качественных признаков A и B , принимающих m_A и m_B значений соответственно. Согласно этому методу, признаки A и B должны быть преобразованы в дихотомические (имеющие лишь два значения):

- признак A' со значениями $a = a_k$ и

$$\bar{a} = \sum_{s \neq k} a_s;$$

- признак B' со значениями $b = b_i$ и

$$\bar{b} = \sum_{s \neq i} b_s.$$

Затем связь между A' и B' измеряется при помощи какой-либо меры связи, представляющей приемлемой для целей исследования.

В качестве такой меры возьмем коэффициент Пирсона χ^2 :

$$\chi^2 = N \left(\sum_{s=1}^{m_A} \sum_{t=1}^{m_B} \frac{(\nu(A_s B_t))^2}{\nu(A_s) \nu(B_t)} - 1 \right), \quad (7)$$

где A_s – множество объектов, соответствующее значению a_s признака A ,
 B_t – множество объектов, соответствующее значению b_t признака B .

Выразим (7) через доли $P_N(\cdot)$:

$$\begin{aligned} \chi^2 &= N \left(\sum_{s=1}^2 \sum_{t=1}^2 \frac{P_N^2(A_s B_t)}{P_N(A_s) P_N(B_t)} - 1 \right) = \\ &= N \left(\frac{P_N^2(AB)}{P_N(A) P_N(B)} + \frac{P_N^2(A \bar{B})}{P_N(A) P_N(\bar{B})} + \right. \\ &\quad \left. + \frac{P_N^2(\bar{A} B)}{P_N(\bar{A}) P_N(B)} + \frac{P_N^2(\bar{A} \bar{B})}{P_N(\bar{A}) P_N(\bar{B})} - 1 \right) = \\ &= N \frac{P_N^2(AB) - 2 P_N(AB) P_N(A) P_N(B) + P_N^2(A) P_N^2(B)}{(1 - P_N(A)) P_N(A) (1 - P_N(B)) P_N(B)} = \\ &= N \frac{(P_N(AB) - P_N(A) P_N(B))^2}{(P_N(A) - P_N^2(A)) (P_N(B) - P_N^2(B))} = N \eta_1(a|b). \end{aligned}$$

Таким образом, мы показали, что

$$\eta_1(a|b) = \frac{\chi^2}{N}. \quad (8)$$

Случай непересекающихся множеств объектов $B_1, \dots, B_m$

Выше предполагалось, что $b_1, \dots, b_m$ могут являться значениями разных признаков $C_1, \dots, C_S$. Пусть теперь все b_i являются значениями какого-то одного признака.

Утверждение 1. Пусть $B_1, \dots, B_m$ – попарно непересекающиеся множества объектов, т.е. $\forall i \neq j \quad B_i \cap B_j = \emptyset$ ($\emptyset$ – пустое множество) и $\nu(B_i B_j) = 0$. Пусть множество $B = \sum_{i=1}^m B_i$; b – значение, которому соответствует множество B . Тогда мера связи

$$\begin{aligned} \eta_m(a|b_1, \dots, b_m) &= \eta_1(a|b) + \\ &+ \frac{1}{P_N(A) - P_N^2(A)} \left(\sum_{i=1}^m \frac{(P_N(AB_i))^2}{P_N(B_i)} - \frac{(P_N(AB))^2}{P_N(B)} \right), \end{aligned}$$

где $P_N(B) = \sum_{i=1}^m P_N(B_i)$ – доля объектов совокупности, принадлежащих B .

Докажем это утверждение, получив в явном виде K_{BB}^{-1} , перемножив матрицы и приведя подобные.

С учетом того, что B_i являются попарно непересекающимися, матрица K_{BB} примет вид

$$K_{BB} = \begin{bmatrix} P_N(B_1) - P_N^2(B_1) & \Lambda & -P_N(B_1) P_N(B_m) \\ -P_N(B_1) P_N(B_2) & \Lambda & -P_N(B_2) P_N(B_m) \\ \text{М} & \text{О} & \text{М} \\ -P_N(B_1) P_N(B_m) & \Lambda & P_N(B_m) - P_N^2(B_m) \end{bmatrix}.$$

Вычислив обратную матрицу, получим

$$K_{BB}^{-1} = \frac{1}{1 - P_N(B)} \begin{bmatrix} h_1 & 1 & \Lambda & 1 \\ 1 & h_2 & \Lambda & 1 \\ \text{М} & \text{М} & \text{О} & \text{М} \\ 1 & 1 & \Lambda & h_m \end{bmatrix},$$

где $h_i = \frac{1 - P_N(B) + P_N(B_i)}{P_N(B_i)}$, $i = \overline{1, m}$. Отсюда

$$\begin{aligned} \eta_m(a) &= \frac{P_N(B)}{(P_N(A) - P_N^2(A)) (P_N(B) - P_N^2(B))} \cdot \\ &\cdot \left\{ \sum_{i=1}^m \sum_{j=1}^m (P_N(AB_i) - P_N(A) P_N(B_i)) (P_N(AB_j) - P_N(A) P_N(B_j)) + \right. \\ &+ \sum_{i=1}^m \frac{1 - P_N(B)}{P_N(B_i)} (P_N(AB_i) - P_N(A) P_N(B_i))^2 \Big\} = \\ &= \frac{(1 - (1 - P_N(B))) (P_N(AB) - P_N(A) P_N(B))^2}{(P_N(A) - P_N^2(A)) (P_N(B) - P_N^2(B))} - \\ &- \frac{P_N(B) (1 - P_N(B))}{(P_N(A) - P_N^2(A)) (P_N(B) - P_N^2(B))} \cdot \\ &\cdot \sum_{i=1}^m \frac{(P_N(AB_i) - P_N(A) P_N(B_i))^2}{P_N(B_i)} = \\ &= \frac{(P_N(AB) - P_N(A) P_N(B))^2}{(P_N(A) - P_N^2(A)) (P_N(B) - P_N^2(B))} + \\ &+ \frac{1}{P_N(A) - P_N^2(A)} \left\{ \sum_{i=1}^m \frac{(P_N(AB_i))^2}{P_N(B_i)} - \frac{(P_N(AB))^2}{P_N(B)} \right\} = \\ &= \eta_1(a|b) + \frac{1}{P_N(A) - P_N^2(A)} \left(\sum_{i=1}^m \frac{(P_N(AB_i))^2}{P_N(B_i)} - \frac{(P_N(AB))^2}{P_N(B)} \right). \end{aligned}$$

Утверждение 2. Если $B_1, \dots, B_m$ – попарно непересекающиеся множества объектов, а $B = \sum_{i=1}^m B_i$, то $\eta_m(a|b_1, \dots, b_m) \geq \eta_1(a|b)$. Причем равенство будет иметь место только в случае

$$\frac{P_N(AB_i)}{P_N(B_i)} = \frac{P_N(AB)}{P_N(B)} \quad \forall i = \overline{1, m} \quad (9)$$

(см. рис. 1). В частности, условие (9) выполняется, когда

- $B = \sum_{i=1}^m B_i \subset A$ (см. рис. 3),
- либо $B = \sum_{i=1}^m B_i \subset \bar{A} = (\Omega \setminus A)$ (см. рис. 4).

В противном случае (когда (9) не выполняется) $\eta_m(a|b_1, \dots, b_m) > \eta_1(a|b)$ (см. рис. 2).

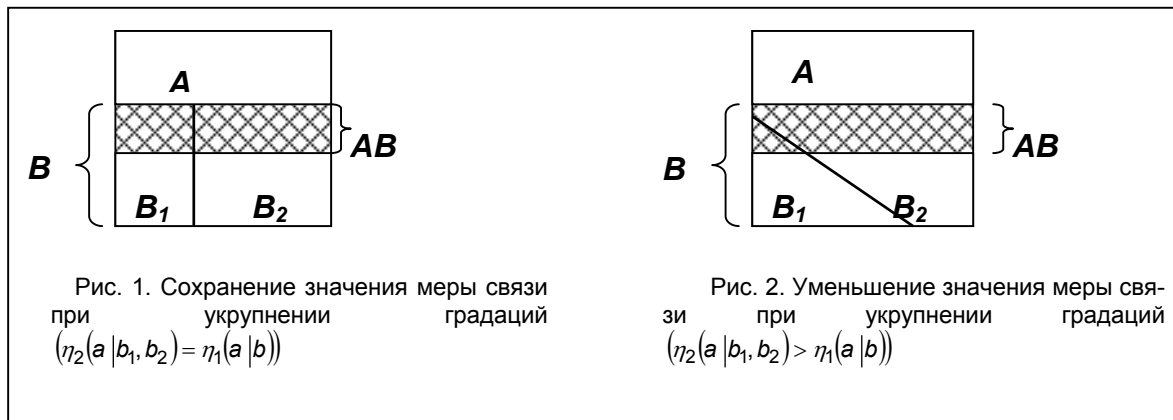
Это утверждение можно доказать, рассмотрев выражение

$$\left(\sum_{i=1}^m \frac{(P_N(AB_i))^2}{P_N(B_i)} - \frac{(P_N(AB))^2}{P_N(B)} \right)$$

как функцию от $P_N(B_i)$ и исследовав его на экстремумы.

$$\eta_1(a|b) = \frac{(0,3 - 0,36)^2}{(0,6)^2 (0,4)^2} \approx 0,34;$$

$$\eta_3(a|b_1, b_2, b_3) \approx 0,34.$$



Данный результат имеет важное практическое значение. Например, с его помощью можно произвести укрупнение градаций признаков. Приведем два примера.

Пример 1. Пусть объектами множества Ω являются люди, принявшие участие в опросе. Обозначим

- A – множество респондентов, имеющих среднемесячный доход свыше 5 тысяч рублей, $P_N(A)=0,6$;
- B_1 – множество респондентов в возрасте от 35 до 40 лет; $P_N(B_1)=0,1$;
- B_2 – множество респондентов в возрасте от 41 до 45 лет, $P_N(B_2)=0,4$;
- B_3 – множество респондентов в возрасте от 46 до 50 лет, $P_N(B_3)=0,1$;
- B – множество респондентов в возрасте от 35 до 50 лет; $P_N(B)=0,6$.

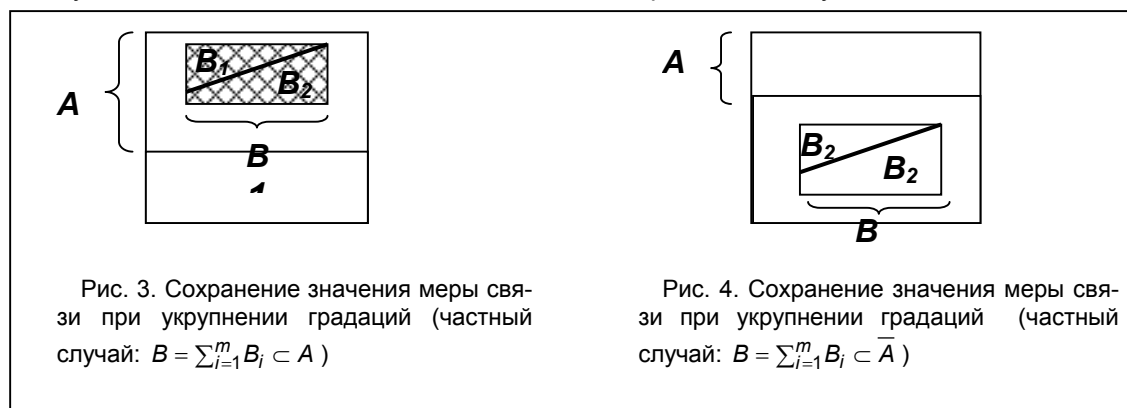
Очевидно $(B_1 + B_2 + B_3) = B$. Пусть $P_N(AB)=0,5$, $P_N(AB_1)=1/12$, $P_N(AB_2)=4/12$, $P_N(AB_3)=1/12$. Вычислив значения меры связи, получим

Таким образом, если основная цель исследования связана с изучением людей с доходом от 5 тыс. рублей, градации b_1, b_2, b_3 признака «возраст» можно объединить в одну градацию b .

Пример 2. Пусть

- множества Ω , A и B – те же, что и в примере 1;
- B_1 – множество респондентов мужского пола в возрасте от 35 до 50 лет; $P_N(B_1)=0,2$,
- B_2 – множество респондентов женского пола в возрасте от 35 до 50 лет, $P_N(B_2)=0,4$;
- B_3 – множество респондентов в возрасте от 35 до 50 лет, имеющих высшее образование, $P_N(B_3)=0,2$;
- B_4 – множество респондентов в возрасте от 35 до 50 лет, не имеющих высшего образования, $P_N(B_4)=0,4$.

Очевидно $(B_1 + B_2) = (B_3 + B_4) = B$. Пусть $P_N(AB)=0,5$, $P_N(AB_1)=1/6$, $P_N(AB_2)=2/6$, $P_N(AB_3)=0,1$, $P_N(AB_4)=0,4$. Вычислив значения меры связи, получим



$$\eta_1(a|b) \approx 0,34;$$

$$\eta_2(a|b_1, b_2) \approx 0,34; \quad \eta_2(a|b_3, b_4) \approx 0,479.$$

Это означает, что если мы разобьем множество B на B_3 и B_4 и посмотрим, как в каждом из них распределились доли респондентов, принадлежащих A , то мы получим больше информации об A , чем дает B в целом. А в случае разбиения B на B_1 и B_2 детализация получается излишней, т.к. при этом мы не получим никакой новой информации об A .

СПИСОК ЛИТЕРАТУРЫ

1. Goodman L.A., Kruskal W.H. Measures of Association for Cross Classifications: I-IV. -Journal of the American Statistical Association, 1954, Vol.49, p.723-764; 1959, Vol.54, p.123-163; 1963, Vol.58, p.310-363; 1972, Vol.67, p.212-241.
2. Пугачев В.Н. Комбинированные методы определения вероятностных характеристик. М.: Советское радио, -1973.
3. Чесноков С.В. Детерминационный анализ социально-экономических данных. М.: Наука, -1982.