УДК 519.23

Об изучении силы связи бинарных показателей

Дронов С.В.
Алтайский государственный университет, г. Барнаул
dsv@math.asu.ru

Аннотация

В работе получено точное распределение коллигативного коэффициента, ранее введенного автором для изучения силы связи между бинарными показателями в качестве альтернативы коэффициенту корреляции Пирсона, применение которого для бинарных показателей не всегда корректно. На основе этого распределения предложен новый статистический критерий, устанавливающий факт связи двух бинарных показателей. Описываются применения этого критерия к методам классификации данных и медицинским задачам дифференциальной диагностики.

Ключевые слова: бинарный показатель, статистическая связь, семейство кластерных разбиений, кластерная метрика.

1. Бинарные показатели в задачах о статистической связи

Задача установления факта связи между изучаемыми показателями, без преувеличения, – центральная тема практически любого научного исследования. Если рассматриваемые показатели являются числовыми, общепринятым инструментом изучения силы их связи является коэффициент корреляции Пирсона. Его неоднократное использование многими исследователями подтвердило интуитивную прозрачность абсолютной величины этого коэффициента как меры силы изучаемой связи, а знака – как ее направления. Практикой многократно были подтверждены выводы, сделанные на основе такой интерпретации.

Тем не менее, хорошо известно, что малые значения коэффициента Пирсона говорят лишь об отсутствии линейной составляющей связи между показателями. В то же время, связи между показателями отнюдь не всегда являются линейными и могут быть устроены гораздо более сложно. Вероятно, особенно ярко это может быть продемонстрировано в случае, когда показатели являются бинарными, т.е. их возможные значения исчерпываются 0 и 1. Действительно, прибегнем к геометрической интерпретации статистической связи с помощью так называемых полей корреляции. Точнее, пусть $X = (x_1, ..., x_n)$, $Y = (y_1, ..., y_n)$ – две связанные выборки. Поставим им в соответствие набор точек $C_i = (x_i, y_i)$, $i = 1, ..., n$. Статистически отличное от нуля значение выборочного коэффициента корреляции $R(X, Y)$ возникает тогда, когда все эти точки группируются к некоторой прямой линии на координатной плоскости. Понятно, что в случае бинарных показателей X, Y имеется всего 4 возможных варианта расположения каждой из точек C_i. Эти точки всегда являются вершинами единичного квадрата, и невозможно построить прямую, близко к которой все они лежат одновременно, за исключением вырожденных случаев.

Если выражаться определеннее, в случае бинарных показателей полностью пропадает возможность говорить о силе статистической связи по величине выборочного коэффициента корреляции. Он здесь не может принимать ни очень малые, ни достаточно большие,

но отличные по модулю от 1, значения. Подобный эффект наблюдается даже в случае, когда бинарным является только один из изучаемых показателей, поэтому здесь оправдано введение в рассмотрение других коэффициентов, описывающих силу связи (см. [1, 2]).

Бинарные показатели на практике встречаются довольно часто, особенно они востребованы при обработке медицинских данных. В частности, именно для оценки степени связи между такими показателями был введен так называемый коэффициент относительного риска RR [3, 4]. Сегодня также активно разрабатываются варианты продвинутых статистических методов для бинарных показателей, см., например, [5, 6].

Ранее, в [7], была, в частности, предпринята попытка ввести новый коэффициент силы связи между показателями бинарного типа. Но, так же, как и при использовании коэффициента относительного риска в медицинских задачах (см. [2,4]), все выводы на основе этого коэффициента могли делаться только чисто эмпирически. Понятия большого и малого коэффициента имели размытые, неточно установленные и строго не обоснованные границы. В настоящей работе этот момент уточнен, и рассматривается строгое определение той величины введенного в [7] коэффициента, при превышении которой можно утверждать, что наличие связи подтверждается статистически значимо.

2. Обозначения. Основная задача работы

Прежде всего подчеркнем, что в данной работе рассматриваются только конечные множества. Если A – такое множество, то через $|A|$ условимся обозначать количество его элементов.

Пусть для каждого из $n \geq 2$ объектов, образующих основное множество $U(|U| = n)$, заданы значения двух бинарных показателей X, Y. Таким образом, в качестве исходных данных выступает пара связанных выборок объема n : $\quad X = (x_1, ..., x_n)$, $Y = (y_1, ..., y_n)$. Следуя [7], введем коэффициент, располагая значением которого на этих выборках, мы сможем сделать вывод о тесноте связи между показателями X, Y. Для этого заметим, что каждая из выборок порождает разбиение множества U на две непересекающиеся части$\mathbf{A} = A_1|A_2$; $\mathbf{B} = B_1|B_2$ так что $U = A_0 \cup A_1 = B_0 \cup B_1$, где

$$A_0 = \{u \in U | X = 0\};\ A_1 = \{u \in U | X = 1\}; \quad B_0 = \{u \in U | Y = 0\};\ B_1 = \{u \in U | Y = 1\}.$$

Эти разбиения будем далее называть разбиениями, индуцированными значениями соответствующих бинарных показателей.

Для определения коэффициента используем кластерную метрику d на семействе всех подобных разбиений, введенную в [8]. Согласно результатам [9], в рассматриваемом случае значения этой метрики можно вычислять по формуле

$$d(\mathbf{A}, \mathbf{B}) = |A_0|^2 + |A_1|^2 + |B_0|^2 + |B_1|^2 - 2\sum_{i=0}^{1}\sum_{j=0}^{1}|A_i \cap B_j|^2. \tag{1}$$

Если через d_* обозначить максимальное значение метрики (1) на всех допустимых парах разбиений **A, B**, то вводимый коэффициент, названный в [7] коллигативным, может вычисляться по формуле

$$K(X, Y) = 1 - \frac{d(\mathbf{A}, \mathbf{B})}{d_*}. \tag{2}$$

Значения коллигативного коэффициента всегда лежат между 0 и 1, и, чем больше его величина, тем более сильной следует признать связь между бинарными показателями.

Очевидно, что множество всех возможных значений $K(X, Y)$ состоит лишь из конечного числа элементов отрезка $[0, 1]$. При этом его значения при переборе всех возможных пар индуцированных показателями разбиений могут совпадать на разных парах. Количество

повторений значения назовем его повторностью. Целью работы является построение статистического критерия значимости $K(X,Y)$. Для этого ниже приводится полное описание множества значений коллигативного коэффициента и указаны значения повторностей элементов этого множества при переборе всех возможных пар индуцированных значениями X, Y разбиений.

3. Теоретические основы критерия

Все необходимое для достижения поставленной цели фактически содержится в следующей элементарной лемме, краткое доказательство которой было дано еще в [7]. Тем не менее, приведем ее несложное доказательство здесь для полноты текста.

Введем обозначения

$$z_{i,j} = |A_i \cap B_j|, \quad i,j = 0,1.$$

Лемма 1. *Формула (1) может быть преобразована к виду*

$$d(\mathbf{A},\mathbf{B}) = d(z) = 2z(n-z), \tag{3}$$

где $z = z_{0,0} + z_{1,1}$.

При этом максимум $d(z)$ *на целых значениях аргумента достигается в точке* $z = [n/2]$ *и равен* $[n^2/2]$ *(*$[t]$ *– целая часть числа* t*).*

Ясно, что если n четное, то точка максимума ровно одна, для нечетного n имеются два одинаковых по величине максимума в точках $z = (n \pm 1)/2$. Все, касающееся максимума функции $d(z)$, – тривиальное следствие свойств квадратичной функции одного переменного, доказывать следует только формулу (3).

Доказательство. Запишем, исходя из (1),

$$d(A,B) = (z_{0,0}+z_{0,1})^2 + (z_{1,0}+z_{1,1})^2 + (z_{0,0}+z_{1,0})^2 + (z_{0,1}+z_{1,1})^2 - 2\sum_{i,j} z_{i,j}^2 =$$
$$= 2(z_{0,0}+z_{1,1})(z_{0,1}+z_{1,0}) = 2(z_{0,0}+z_{1,1})(n - z_{0,0} - z_{1,1}).$$

Этим доказательство леммы завершается. □

Из доказанной леммы несложно выводится следующее утверждение.

Лемма 2. *Все значения* $d(\mathbf{A},\mathbf{B})$ *содержатся в множестве целых чисел* $\{0,...,[n^2/2]\}$*, но не каждый элемент этого множества служит ее значением. Значение 0 достигается только на паре совпадающих разбиений* $\boldsymbol{A}$ *и* $\boldsymbol{B}$*, а* $[n^2/2]$ *заведомо достигается на некоторой паре разбиений.*

На самом деле все значения метрики в этом интервале исчерпываются числами, равными удвоенной площади прямоугольника периметра $2n$, имеющего стороны, длины которых выражаются целыми числами. При этом из (3) вытекает, что каждое из своих значений, кроме, возможно, максимального, $d(z)$ принимает при двух различных z. Это значение z станет единственным, если потребовать $z \le n/2$. При этом дополнительном условии функция d становится обратимой.

Теорема 1. *Пусть* d_0 *– одно из возможных значений метрики* d *и* $s = d^{-1}(d_0)$*, т.е.*

$$2s(n-s) = d_0, \quad s \le n/2, \quad Q(s) = 2^{n-1}C_n^s.$$

Тогда при $s < \frac{n}{2}$ *имеется всего* $Q(s)$ *упорядоченных пар разбиений основного множества, таких, что* $d(\mathbf{A},\mathbf{B}) = d_0$*. Если же* n *четное, и* $s = \frac{n}{2}$*, то таких пар имеется* $\frac{Q(n/2)}{2}$.

Доказательство. Очевидно, что для получения значения метрики, равного d_0, необходимо и достаточно построить такие разбиения $\mathbf{A} = A_1|A_2$; $\mathbf{B} = B_1|B_2$, чтобы во введенных при доказательстве леммы 1 обозначениях $z_{0,0} + z_{1,1} = s$. Будем перебирать все такие возможности следующим образом: сначала выберем $V \subset U$, состоящее из s элементов (C_n^s способами), а затем будем строить 4 дизъюнктных множества $V_{i,j}$, $i, j = 0, 1$ так, чтобы

$$V_{0,0} \cup V_{1,1} = V, \quad V_{1,0} \cup V_{0,1} = U \backslash V.$$

Примем число элементов $V_{i,j}$ за $z_{i,j}$ при произвольных i, j. Тогда все нужные нам разбиения имеют при некотором конкретном выборе введенных множеств вид

$$A_0 = V_{0,0} \cup V_{0,1},\ A_1 = V_{1,1} \cup V_{1,0}; \quad B_0 = V_{0,0} \cup V_{1,0},\ B_1 = V_{1,1} \cup V_{0,1}.$$

При этом разбиения $\mathbf{A}$ и $\mathbf{B}$ можно поменять местами, но неупорядоченная пара разбиений при таком выборе не изменится. Заметим, что при этом каждая упорядоченная пара разбиений образуется ровно один раз, если $s \neq \frac{n}{2}$.При четном n и $s = \frac{n}{2}$ за счет возможности выбора в качестве V второй из частей множества, которая состоит также из $n/2$ элементов, каждая такая пара разбиений появится в описанном построении дважды.

Всего множество V можно разбить на две непересекающиеся части 2^s способами, кодируя каждую из частей 0 или 1. Но при этом перемена местами этих символов приведет к тому же разбиению, поэтому всего разных способов построить $V_{0,0}$ и $V_{1,1}$ имеется 2^{s-1}. Аналогично, $V_{0,1}$, $V_{1,0}$ можно построить 2^{n-s-1} способами. Тогда общее число способов получить требуемое значение метрики на упорядоченной паре разбиений равно ($s \neq n/2$)

$$2 \cdot C_n^s \cdot 2^{s-1} \cdot 2^{n-s-1} = Q(s).$$

Если же $s = n/2$, то это значение следует разделить на два, поскольку каждая упорядоченная пара разбиений здесь повторяется. Теорема доказана. □

Предположим теперь, что каждое из двух разбиений генерируется случайно. Поскольку, как уже было отмечено, всего множество из n элементов может быть разбито на две дизъюнктные части 2^{n-1} способами, то каждая конкретная упорядоченная пара различных разбиений будет сгенерирована с вероятностью

$$p_{\mathbf{A},\mathbf{B}} = \frac{1}{\left(2^{n-1}\right)^2} = \frac{1}{2^{2n-2}},$$

а если $\mathbf{A} = \mathbf{B}$, то соответствующая вероятность окажется вдвое меньшей.

Привлекая утверждение только что доказанной теоремы, немедленно получаем

Следствие 1. *Пусть разбиения $\boldsymbol{A}$, $\boldsymbol{B}$ основного множества U на две дизъюнктные части генерируются случайно и независимо друг от друга. Пусть t – допустимое значение метрики, $s = d^{-1}(t)$ Тогда*

$$P(d(A, B) = t) = \frac{C_n^s}{2^{n-1}},\ t \neq \left[\frac{n^2}{2}\right] \tag{4}$$

и вдвое меньше при $t = [n^2/2]$.

Таким образом, распределение метрики d получается как бы “склеиванием” значений биномиального распределения, симметричных относительно середины их ряда.

4. Статистический критерий значимости коллигативного коэффициента

Из общего определения (2) коэффициента и леммы 1 вытекает, что для оценивания силы связи бинарных показателей коллигативный коэффициент равен

$$K(X,Y) = 1 - \frac{d(\mathbf{A},\mathbf{B})}{[n^2/2]}. \tag{5}$$

Числитель дроби (5) удобно вычислять по формуле (3). Напомним, что в (3) $z_{0,0}$ – число элементов основного множества, на которых оба показателя принимают значение 0, а на $z_{1,1}$ элементах этого множества оба они равны 1.

Таким образом, вероятность того, что при имеющимся конкретном значении этого коэффициента K связь между порождающими соответствующие разбиения бинарными показателями отсутствует, равна

$$p(K) = \sum_{t\,>\,[n^2/2](1-K)} P(d(\mathbf{A},\mathbf{B}) = t) = \frac{1}{2^{n-1}} \sum_{t\,>\,[n^2/2](1-K)} C_n^{d^{-1}(t)}, \tag{6}$$

и следует признать связь между показателями статистически значимой, если эта вероятность достаточно мала, например, меньше стандартной величины $p = 0,05$.

Рассмотрим пример, взятый из реальной медицинской практики [10]. Ниже приведена таблица 1, в которой для каждого из 12 пациентов в первой строке указано наличие патологического гена PAI-1 (ингибитор активатора плазминогена), а во второй – признак наличия тромбоза глубоких вен.

Расчеты по формулам (1) и (5) приводят к следующим результатам

$$d(\mathbf{A},\mathbf{B}) = 54, \quad K = 0,25.$$

Таблица 1

К связи генотипа с тромбозом

PAI-1	1	1	1	1	0	0
ТЭЛА	1	1	1	1	1	1
PAI-1	0	0	1	1	1	1
ТЭЛА	0	0	0	0	0	0

Для оценки значимости силы связи воспользуемся (6) и таблицей 2. Эта таблица построена с помощью формулы (4). В первой строке ее собраны все возможные значения метрики d в этом случае, во второй – соответствующие им значения z из формулы (3), в третьей – значения коллигативного коэффициента для величины d в (5).

Таблица 2

Ряд распределения метрики и коллигативного коэффициента для $n = 12$

d	0	22	40	54	64	70	72
z	0	1	2	3	4	5	6
K	1	0,6944	0,4444	0,2500	0,1111	0,0278	0
вероятность	0,0005	0,0059	0,0322	0,1074	0,2417	0,3867	0,2256

Вычисления дают

$$p(K) = 0,0005 + 0,0059 + 0,0322 = 0,0386,$$

что на стандартном уровне $p = 0,05$ означает наличие статистически значимой связи между наличием соответствующего паталогического гена и возникновением тромбоза.

Заметим, что применение обычной в этом случае методики к данным таблицы 1 дает значение выборочного коэффициента корреляции Пирсона 0,192, что приводит к его уровню значимости $p = 0,274$, т.е. к выводу об отсутствии значимой корреляционной связи между этими показателями. Врачебная практика, однако, наличие связи все же подтверждает. Таким образом, вновь предлагаемый подход к оцениванию силы связи между бинарными показателями, по крайней мере, на рассматриваемом примере, в большей степени соответствует практике медицинской диагностики.

5. Выводы и заключение

В работе предложен новый подход к оцениванию силы связи между бинарными показателями. Он основан на изучении степени различия двух разбиений основного множества, индуцированных значениями этих показателей. Связь считается тем более сильной, чем в меньшей степени различаются эти разбиения. Близость разбиений, а следовательно, и степень связи предлагается описывать специальным коллигативным коэффициентом, принимающим значения в интервале между 0 и 1. На медицинском примере показано, что предлагаемый способ дает более согласующиеся с практикой результаты, чем обычный корреляционный анализ.

Полученное в работе точное распределение коллигативного коэффициента позволяет делать статистически достоверные выводы не только о наличии связи между бинарными показателями, но и о различии двух произвольных разбиений изучаемого множества объектов на две дизъюнктные части. Это дает возможность, например, сравнивать разные методы классификации одних и тех же объектов, что весьма востребовано, скажем, при введении в медицинскую практику новых методов дифференциальной диагностики и в иных задачах доказательной медицины (по этому поводу подробнее в [11]).

Список литературы

1. Дронов С.В., Бойко И.Ю. Метод оценки степени связи бинарного и номинального показателей // Прикладная дискретная математика. — 2015. — № 4. — С. 109–119.

2. Дронов С.В., Шепелев С.А. Сравнение подходов к оценке степени связи нечисловых факторов в четырехпольных таблицах // Известия Алтайского государственного университета. — 2014. — № 1/2. — С. 31–34.

3. Sistrom C.L., Garvan C.W. Proportions, odds and risk // Radiology. — 2004. — Vol. 230, v.(1). — P. 12–19.

4. Cummings P. The relative merits of risk ratios and odds ratios // Arch Pediatr. Adolesc. Med. — 2009. — Vol. 163(5). — P. 438–445.

5. Frolov A.A., Sirota A.M., Husek D. et al. Binary factorization in Hopfield-like neural networks: Single-step approximation and computer simulations // Neural Network World. — 2009. — Vol. 14(2). — P. 139–152.

6. Keprt A., Snasel V. Binary Factor Analysis with Genetic Algorithms // Soft Computing as Transdisciplinary Science and Technology, Proceedings of the fourth IEEE International Workshop WSTST'05. — Springer, 2005. — P. 1259–1268.

7. Дронов С.В., Фоменко А.П. Способ оценки прогностической силы бинарного показателя // Известия АлтГУ. — 2017. — № 4. — С. 89–93.

8. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. — 2011. — № 1/2. — С. 32–35.

9. Dronov S.V., Eudokimov E.A. Post-hoc cluster analysis of connection between forming characteristics // Model Assisted Statistics and Applications. — 2018. — Vol. 13, no. 2. — P. 183–192.

10. Петриков А.С., Шойхет Я.Н., Белых В.И и др. Молекулярно-генетические основы развития гипергомоцистеинемии у больных с венозными тромбоэмболическими осложнениями // Медицина и образование в Сибири (сетевое научное издание). — 2013. — № 2. — www.ngmu.ru/cozo/mos/article/text_full.php?id=997.

11. Straus Sh.E., McAlister F.A. Evidence-based medicine: a commentary on common criticisms // Canadian Medical Association Journal. — 2000. — Vol. 163, no. 7. — P. 837–841.