УДК 519.23

Способ оценки прогностической силы бинарного показателя

С.В. Дронов, А.П. Фоменко
Алтайский государственный университет (Барнаул, Россия)

A Method for Estimating the Predictive Power of a Binary Indicator

S.V. Dronov, A.P. Fomenko
Altai State University (Barnaul, Russia)

В работе рассматривается решение одной из разновидностей задач классификации данных. Допустим, все рассматриваемое множество объектов разбито каким-то образом на две группы (правильное разбиение). Наряду с этим у каждого из рассматриваемых объектом измерен некоторый бинарный показатель – у каждого из объектов он принимает значение 0 или 1. Требуется оценить, насколько уверенно знание этого показателя позволяет отнести объект к одной из групп правильного разбиения. Такого рода задача является разновидностью задачи дискриминантного анализа, где правило, относящее объект к одной из групп, называют прогностическим. Поэтому вводимая в работе числовая характеристика степени информативности показателя названа прогностической силой бинарного показателя. Она вводится путем оценки различий правильного разбиения множества и разбиения, построенного по изучаемому бинарному показателю. Величина различия определяется путем расчета кластерной метрики, ранее введенной в работе первого автора. Производится сравнение этой характеристики с традиционно используемыми в этом случае коэффициентами корреляции и коэффициентом относительного риска.

Ключевые слова: классификация данных, прогностическое правило, бинарный показатель, кластерная метрика.

The paper considers the solution of one of the types of classification problems in data analysis. Let us assume that the set of objects under consideration is in some way divided into two groups (we call this a regular partition). Along with this, each of the objects in view has a certain binary indicator measured – for each of the objects it has only values 0 or 1. It is required to estimate the confidence to assign the object to one of the groups of the regular partition using the knowledge of this indicator. This problem is a variant of the so called discriminant analysis problem, where the rule for assigning an object to one of the groups is called prognostic. So, the introduced numerical characteristic of the indicator of the informational content is called the prognostic power of it. The characteristic is introduced by estimating the differences between the regular partition of the set and the partition constructed by the binary indicator being studied. The magnitude of the difference is determined by calculating the cluster metric previously introduced in the work of the first author. This characteristic is compared with the correlation coefficient and the relative risk ratio commonly used in such cases.

Key words: data classification, prognostic rule, binary variable, cluster metric.

DOI 10.14258/izvasu(2017)4-15

1. Обоснование и постановка задачи. Проблема классификации объектов исследования возникает в любой области науки. Инструменты для решения этой задачи весьма разнообразны и достаточно хорошо разработаны (см., например, до сих пор актуальную книгу [1], целиком посвященную этой теме, а также [2]). В частности, когда классы объектов, подлежащих изучению, должным образом определены и описаны, эта проблема сводится к выбору того из имеющихся классов, к которому должен быть отнесен вновь обнаруженный объект, и получила название дискриминантного анализа, методы которого сегодня активно развиваются (см. [3]). Одно из востребованных применений этого аппарата – задачи так называемой доказательной медицины, основы которой изложены в [4, 5]. Речь здесь идет, например, о проблеме дифференциальной диагностики или создании математически обоснованных правил, называемых прогностическими, которые

по данным медицинских исследований позволяют поставить пациенту тот или иной диагноз. При этом часто данные медицинских исследований представлены не в числовой, а в бинарной форме (наблюдается определенный синдром или нет, были ли пройдены определенные лечебные процедуры и т.п.).

При построении прогностических правил традиционными способами, как правило, предполагается, что данные, по которым они строятся, имеют числовой характер, непрерывную шкалу значений. В теоретических обоснованиях эффективности подобных правил (например, [6]) дополнительно вводится предположение, что показатели исследуемых объектов имеют нормальные распределения, что заведомо неверно в предположении их бинарности. Поэтому бинарная ситуация требует отдельного изучения.

Предположим, что исследуемое множество U объектов разбито на два непустых подмножества A_1 и A_2 (здоровые и больные пациенты). Это разбиение будем обозначать $\mathcal{A} = A_1!A_2$ и называть правильным разбиением. Наряду с этим у каждого из объектов имеется некий бинарный показатель Z. Основная задача работы – ввести числовую характеристику, оценивающую его прогностическую силу, т. е. уровень доверия к возможности определять, в какое из подмножеств правильного разбиения следует отнести объект, зная только значение этого Z.

Для введения требуемой характеристики предлагается построить новое разбиение $\mathcal{B} = B_1!B_2$ того же множества U, относя в одно из подмножеств этого разбиения те объекты, для которых $Z = 1$, а в другое те, для которых $Z = 0$. Назовем его разбиением по Z и сравним два полученных разбиения. Прогностическая сила Z должна быть тем больше, чем в большей степени схожи эти разбиения – правильное и по Z.

Использовать введенную характеристику на практике возможно, например, тогда, когда установлена высокая прогностическая сила Z, а наличие этого синдрома, т.е. условие $Z = 1$, проверить оказывается проще, чем применять традиционные методы диагностики.

Следует также отметить, что решаемая задача по своей сути близка к так называемой post-hoc-задаче кластерного анализа, описание которой можно найти, например, в [7], но в силу бинарности рассматриваемых показателей не совпадает с ней в ее традиционной постановке.

2. Кластерная метрика и ее диапазон на семействе 2-разбиений. В качестве меры различия двух разбиений одного и того же конечного множества на непустые подмножества (далее будем употреблять термин разбиение) будем использовать кластерную метрику, которая была введена в [8]. Она определяется следующим образом. Условимся через $|A|$ обозначать количество элементов конечного множества A. Рассмотрим два разбиения $\mathcal{A}, \mathcal{B}$ основного множества U. Для каждого $x \in U$ найдутся множества A_x, B_x, его содержащие и являющиеся элементами первого и второго разбиений соответственно. Тогда величина кластерной метрики по определению равна

$$d(\mathcal{A}\,, \mathcal{B}) \;= \sum_{x \in U} |A_x \Delta B_x|,$$

где $A_x \Delta B_x = (A_x \setminus B_x) \cup (B_x \setminus A_x)$ – симметрическая разность множеств.

В цитированной работе была доказана формула, позволяющая вычислять введенную метрику более удобным способом. Конкретизируем состав рассматриваемых разбиений. Пусть $\mathcal{A} = \{A_1, ..., A_s\}$, $\mathcal{B} = \{B_1, ..., B_t\}$. Введем обозначения

$$\cap_{i,j} = |A_i \cap B_j|, \quad T_{i,j} = |A_i \Delta B_j|.$$

Тогда справедливо

$$d(\mathcal{A}\,, \mathcal{B}) \;=\; \sum_{i,j} \cap_{i,j} T_{i,j}. \qquad (1)$$

В [8] доказано также, что максимально возможное значение метрики на семействе всех разбиений множества из n элементов равно $n(n-1)$ и достигается оно тогда и только тогда, когда $s = n, t = 1$, или наоборот.

Вернемся к нашей задаче. Мы рассматриваем лишь разбиения, состоящие из двух подмножеств. Будем называть их 2-разбиениями. В этом случае понятно, что, если $|U| = n$, то значение $d = n(n-1)$ не может быть достигнуто. Займемся поиском достижимого максимума в рамках решаемой задачи.

Лемма 1 (основная лемма). *Пусть заданы два 2-разбиения* $\mathcal{A} = A_1!A_2, \mathcal{B} = B_1!B_2$ *и* $s = |A_1 \Delta B_1|$. *Тогда*

$$d(\mathcal{A}\,, \mathcal{B}) \;=\; 2s(n-s). \qquad (2)$$

Доказательство. Пусть $W_{i,j} = A_i \cap B_j$. Тогда $|W_{i,j}| = \cap_{i,j}$ во введенных обозначениях. Условимся далее "двойственное значение" индекса i обозначать i': $i' = 3 - i$. Заметим тогда, что

$$A_i \Delta B_j = W_{i,j'} \cup W_{i',j}, \quad \overline{W_{i,j}} \;=\; W_{i,j'} \cup W_{i',j} \cup W_{i',j'},$$

а также $s = \cap_{1,2} + \cap_{2,1}$. Следовательно,

$$A_{i'} \Delta B_{j'} = W_{i',j} \cup W_{i,j'} = A_i \Delta B_j$$

и, кроме этого,

$$\overline{A_i \Delta B_j} \;=\; \overline{W_{i,j'}} \cap \overline{W_{i',j}} = W_{i,j} \cup W_{i',j'} = A_i \Delta B_{j'}.$$

Таким образом, из четырех симметрических разностей множеств первого и второго разбиений имеется всего две различных, причем, например,

$A_1\Delta B_1, A_2\Delta B_2$ одинаковы, а количества элементов $A_1\Delta B_1$ и $A_1\Delta B_2$ в сумме дают n. Применим формулу (1):

$$d(\mathcal{A},\mathcal{B}) = \cap_{1,1}s + \cap_{1,2}(n-s) + \cap_{2,1}s + \\ + \cap_{2,2}(n-s) = s(\cap_{1,1}+\cap_{2,2}) + \\ +(n-s)(\cap_{1,2}+\cap_{2,1}) \;=\; 2s(n-s).$$

Лемма доказана.

Заметим, что (2) не меняется при замене s на $n-s$. Поэтому без ограничения общности можно считать, что множества в разбиениях пронумерованы таким образом, что $s = |A_1\Delta B_1| \le [n/2]$ ($[n/2]$ – целая часть числа $n/2$).

Лемма 2. *Для любого $s \in \{1,2,...,[n/2]\}$ и произвольного 2-разбиения $\mathcal{A} = A_1!A_2$ существует 2-разбиение $\mathcal{B} = B_1!B_2$ такое, что величина $d(\mathcal{A},\mathcal{B})$ задается (2).*

Доказательство. Пусть множество A_1 содержит не меньше элементов, чем A_2. Тогда $s-1 < |A_1|$. Выберем

$$C = \{x_1,...,x_{s-1}\} \subset A_1, \quad x_s \in A_2$$

произвольно. Положим $B_1 = (A_1 \setminus C)\cup\{x_s\}$, $B_2 = U \setminus B_1$. Тогда $A_1\Delta B_1 = \{x_1,...,x_s\}$, и, согласно лемме 1, $d(\mathcal{A},\mathcal{B}) = 2s(n-s)$. Лемма доказана.

Теорема 1. *Все возможные значения метрики d на семействе 2-разбиений множества U из n элементов исчерпываются возрастающей цепочкой чисел d_j, $j = 1,2,...,[n/2]$,*
где

$$d_1 = 2(n-1)$$
$$d_{s+1} = d_s + 2(n-2s-1),\ s = 1,...,[n/2]-1;$$
$$d_{[n/2]} = [n^2/2].$$

Доказательство. Рассмотрим $d_s = 2s(n-s)$, где $s \in \{1,2,...,[n/2]\}$. Согласно доказанным леммам 1, 2 иных значений метрика принимать не может, и каждое такое значение соответствует некоторой паре 2-разбиений U. Выписанные в утверждении теоремы соотношения легко проверяются непосредственно. Теорема доказана.

3. Прогностическая сила бинарного показателя. Заметим, что, если d фиксировано, то для тех 2-разбиений $\mathcal{B} = B_1!B_2$, которые удалены от данного $\mathcal{A} = A_1!A_2$ на d, число $s = |A_1\Delta B_1|$ определяется из формулы (2) как

$$s \;=\; s(d) = \frac{1}{2}\left(n - \sqrt{n^2-2d}\right). \qquad (3)$$

Знак "$-$" перед радикалом здесь может быть выбран без ограничения общности в силу замечания после основной леммы. Поэтому всегда $s = s(d) \le n/2$. Преимущество перехода от d к $s(d)$ состоит еще и в том, что при увеличении d к следующему его возможному значению из теоремы, s просто увеличивается на один, тогда как соседние значения d вычисляются сложнее.

Теорема 2. *Пусть в 2-разбиении $\mathcal{A} = A_1!A_2$ $|A_1| = k \le n/2$. Тогда количество 2-разбиений этого множества, удаленных от него на величину d из допустимого диапазона, указанного в теореме 1, равно*

$$m(d) \;=\; m(d(s)) = \; C_n^s - \chi_1(k,s) - \chi_2(k,s),$$

где s задается формулой (3), $\chi_1(k,s) = 1$, если $s = k \ne n/2$, иначе 0, а $\chi_2(k,s) = C_n^{n/2}/2+1$, если n четное, $s = k = n/2$, иначе 0.

Доказательство. Построить требуемое разбиение в силу основной леммы возможно путем построения такого $B_1 \subset U$, чтобы $|A_1\Delta B_1| = s$. Выберем подмножество $C \subset U$, $|C| = s$ произвольно. Положим

$$B_1 \;=\; A_1\Delta C, \quad B_2 \;= \overline{B_1} \;=\; A_1\Delta\overline{C}. \qquad (4)$$

Нетрудно проверить, что $A_1\Delta B_1 = C$, а значит, требуемое разбиение построено. Таким образом, чтобы построить все такие разбиения, следует перебрать все подмножества U из s элементов и применить (4), что можно сделать C_n^s способами. Описанное построение не даст разбиения только в том случае, когда $C = A_1$ – здесь B_1 получится пустым, или $C = A_2$, что приведет к $B_1 = U$, а следовательно, к $B_2 = \emptyset$. Первая ситуация возможна только тогда, когда $s = k$, и этот единственный способ следует исключить из общего числа. Вторая же ситуация невозможна для $k, s < n/2$, и только в случае, когда в $\mathcal{A}$ оба подмножества одинаковы по величине, подлежат исключению два варианта ($C = A_1$ или $C = A_2$).

Повторяться построенные разбиения могут только в случае, когда в качестве B_2 нового разбиения получится B_1 одного из построенных ранее разбиений. Заметим, что в силу предложенного алгоритма построения это означало бы, что нашлись бы множества C, C_1, оба содержащие по s элементов, такие, что

$$A_1\Delta C = \overline{A_1\Delta C_1} = A_1\Delta\overline{C_1}.$$

Но это равенство возможно лишь для $C_1 = \overline{C}$. Поэтому повторения возникают только при четных n, $s = n/2$, и при этом каждое разбиение окажется учтенным точно два раза. Таким образом, в этом случае число разбиений будет равно $C_n^{n/2}/2$, и, как и в предыдущем случае, следует исключить одно разбиение, при котором одно из множеств окажется пустым. Теорема доказана.

Лемма 3. *Всего существует $2^{n-1}-1$ различных 2-разбиений множества из n элементов.*

Доказательство. Любое 2-разбиение соответствует цепочке из цифр 0 и 1 длины n, написанных напротив каждого из элементов U: те элементы U, напротив которых будет написана 1, отнесем в первое множество разбиения, остальные –

во второе. Поскольку пустые множества в разбиении недопустимы, следует запретить цепочки, состоящие из одинаковых цифр. Получим $2^n - 2$ вариантов. При этом, если мы все нули поменяем на единицы и наоборот, то разбиение не изменится. Поэтому на самом деле их будет вдвое меньше. Лемма доказана.

Теперь вернемся к основной задаче. Пусть $\mathcal{A} = A_1!A_2$ – правильное 2-разбиение изучаемого множества объектов U. Если предположить, что статистически значимой связи между некоторым бинарным показателем Z и этим разбиением нет, то разбиение, построенное по Z, может оказаться любым из $2^{n-1} - 1$ возможных 2-разбиений множества U с равными вероятностями. Учитывая это, можно предложить следующий статистический критерий проверки значимости прогностической силы бинарного показателя Z.

Пусть $|A_1| = k$. Определим s как сумму числа элементов в A_1, для которых $Z = 0$ и числа элементов A_2, для которых $Z = 1$ ($s = |A_1 \Delta B_1|$). Если окажется, что $s > n/2$, то изменим s, вычтя это большое значение из n: $s := n - s$. Найдем $d = 2s(n - s)$. Выберем достаточно малое $\varepsilon > 0$.

Используя формулы теоремы 3 при найденных параметрах, вычислим

$$N(d) \;=\; \sum_{j=s}^{[n/2]} m(d(j)).$$

Если $\frac{N(d)}{2^{n-1}-1}$ близко к 1, то прогностическую силу Z следует признать статистически значимой. Точнее, если оно больше $1 - \varepsilon$, то, в случае отсутствия связи Z с правильным разбиением, получение настолько же или более близкого к правильному разбиению по Z было бы практически невероятно (имело бы вероятность, меньшую ε).

4. Обсуждение и выводы. Пусть описанным выше способом установлена статистически значимая прогностическая сила бинарного показателя Z. Для использования этого показателя при практической диагностике следует разобраться, какое из значений (0 или 1) этого показателя должно сопровождаться отнесением объекта в первое множество правильного разбиения. Будем считать, что первые множества обоих разбиений соответствует значениям 1 показателей. Если $s = |A_1 \Delta B_1|$ не больше $n/2$, то таким значением будет $Z = 1$, иначе $Z = 0$.

Заметим, что признание наличия значимой прогностической силы означает признание существования значимой статистической связи между бинарными показателями Z и Y, который определяет правильное разбиение. Поэтому, введя обозначение для числа прогностической силы

$$J(Y, Z) \;=\; \frac{N(d)}{2^{n-1} - 1},$$

можно попробовать произвести сравнение предложенного метода с методами, использующими проверку значимости разного рода характеристик связи между этими показателями. В частности, наиболее распространенными показателями являются коэффициент корреляции Пирсона (или его вариант – коэффициент бисериальной корреляции), а также так называемый коэффициент относительного риска. Он для двух бинарных показателей определяется следующим образом. По значениям пар (Y, Z) вычисляют четыре числа: a – количество пар (1, 1), b – пар (1,0) и c, d пар (0,1) и (0,0) соответственно. Коэффициент относительного риска полагают равным

$$RR \;=\; \frac{a}{a+b} \cdot \frac{c+d}{c}$$

и считают факт связи между Y, Z установленным, если RR значимо отличен от 1. Этот коэффициент широко применяют в медицине (многочисленные примеры есть в [9]), математические исследование и обоснование свойств этого коэффициента можно найти в [10].

Например, пусть правильное разбиение множества из 8 элементов задано значениями Y во втором (и шестом) столбце таблицы, а другие два разбиения (по синдромам Z, T) заданы столбцами 3, 7 и 4, 8 этой таблицы.

Три разбиения

Объект	Y	Z	T	Объект	Y	Z	T
1	1	0	0	5	1	0	1
2	1	1	0	6	1	1	1
3	1	1	1	7	1	1	1
4	0	0	1	8	0	0	0

При расчете характеристик связи между Y и Z получаем $J(Y, Z) = 0,94$, коэффициент корреляции $\rho(Y, Z) = 0,58$, а RR принимает бесконечно большое значение. Все характеристики подтверждают наличие связи между показателями. Это правильно, поскольку по-разному здесь классифицируются только первый и пятый объекты. С другой стороны, $J(Y, T) = 0,71, \rho(Y, T) = 0,15, RR(Y, T) = 1,33$. Здесь число прогностической силы указывает на наличие связи между показателями, в то время как остальные характеристики этого не подтверждают. Таким образом, можно сказать, что прогностическая сила бинарного показателя выявляет новый вид связи, не совпадающий с ранее изучавшимися. Тем не менее, если два рассматриваемых разбиения совпадают, результаты оценки степени связи таких бинарных показателей окажутся одинаковыми – $J(X, Y)$ будет равен 1, коэффициент корреляции окажется равным ± 1, а коэффициент относительного риска окажется либо 0, либо примет бесконечно большое значение.

Библиографический список

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М., 1989.

2. Mills P. Efficient statistical classification of satellite measurements // International Journal of Remote Sensing. – 2011. – № 32(21). DOI: 10.1080/01431161.2010.507795.

3. Haghighat M., Abdel-Mottaleb M. & Alhalab W. Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition. // IEEE Transactions on Information Forensics and Security. – 2016. – V. 11, № 9. DOI: 10.1109/TIFS.2016.2569061.

4. Straus S., Glasziou P., Scott Richardson W., Brian Haynes R. Evidence Based Medicine. – Elsevier, Churchhill, Livingstone, 2010.

5. Sackett D.L Rosenberg W.M. Gray J.A. Haynes R.B. Richardson W.S. Evidence based medicine: what it is and what it isn't. // BMJ. – 1996. – № 312 (7023). DOI:10.1136/bmj.312.7023.71.

6. McLachlan G. Discriminant Analysis and Statistical Pattern Recognition. – Wiley, 2004.

7. Дронов С.В. Методы и задачи многомерной статистики. – Барнаул, 2015.

8. Dronov S.V., Dementjeva E.A. A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. – 2012. – Vol. 7, № 1. DOI: 10.3233/MAS-2011-02-01.

9. Crawford-Brown D.J. Theoretical and Mathematical Foundations of Human Health Risk Analysis: Biophysical Theory of Environmental Health Science. – Springer Science & Business Media, 2012.

10. Fleiss J., Levin B. Statistical Methods for Rates and Proportions. – Wiley, 2003.