УДК 517

ОЦЕНКА НЕСЛУЧАЙНОСТИ ИНТЕРВАЛЬНЫХ ИМПЛИКАТИВНЫХ ЗАВИСИМОСТЕЙ, СОДЕРЖАЩИХСЯ В БУЛЕВЫХ ТАБЛИЦАХ ДАННЫХ

Ильченко А.В.

Таврический национальный университет им. В.И. Вернадского,
факультет математики и информатики
пр-т Вернадского, 4, г.Симферополь, Крым, Украина, 95007

Abstract

An interval implicative dependency concept and a statistical estimation of its nonrandom appearance in Boolean empirical data tables are considered in the paper.

Введение

Благодаря легкости интерпретации и наглядности получаемых результатов, методы поиска логических закономерностей в экспериментальных данных приобретают все большую популярность в анализе данных. Обычно реальные данные содержат большое число точных и приближенных ассоциативных правил, среди которых много избыточных, случайных и малозначимых. И, если устранение избыточных правил решается с помощью алгоритмов, которые строят базис (в том или ином смысле [2], [3]) ассоциативных правил, то вопросы обоснованности, объективности, неслучайности получаемых результатов, остаются во многом открытыми и актуальными.

Целью данной работы является получение статистической оценки неслучайности возникновения интервальной импликативной зависимости в булевых таблицах эмпирических данных.

Инрервальная импликативная зависимость.

Пусть $B^n = \{0,1\}^n$ - булево признаковое пространство.
G - множество объектов.
H - проекция множества G в пространство B^n («занятая» область B^n).
$\overline{H} = B^n \backslash H$ - дополнение проекции H до B^n («свободная» область B^n).
$I(B^n)$ - обозначение множества всех интервалов пространства B^n.

Определение 1. Пусть $I, J \in I(B^n)$ - интервалы булевого пространства $B^n, H \subseteq B^n$. На подмножестве H точек пространства B^n выполняется интервальная импликация $I \to J$, если все точки подмножества H, содержащиеся в интервале I, а также содержатся в интервале J:

$$I \to J, \quad \text{если} \quad I \cap H \subseteq J \cap H. \tag{1}$$

Интервал I - посылка, интервал J- заключение импликации $I \to J$. Ранги интервалов I и J - ранги посылки и заключения соответственно.

Если импликация $I \to J$ выполняется, то все точки подмножества H попавшие в интервал I, содержатся в пересечении интервалов I и J, то есть

$$I \cap J \subseteq (I \cap J) \cap H. \tag{2}$$

Область $I \backslash J = I \backslash (I \cap J)$ не содержит точек подмножества H. Именно наличие таких областей, свободных от точек подмножества H, служит причиной появления импликативных зависимостей в пространстве B^n.

Теорема 1. *Пусть $I, J \in (B^n)$ - интервалы булевого пространства B^n, $H \subseteq B^n$. Импликация $I \to J$ выполняется на подмножестве H точек пространства B^n тогда и только тогда, когда область $I \backslash J$ не пересекается с подмножеством H:*

$$I \to J \Leftrightarrow (I \backslash J) \cap H = \varnothing. \tag{3}$$

Доказательство. Необходимость. Пусть импликация $I \to J$ выполняется на подмножестве H точек пространства B^n. Допустим, что $(I \backslash J) \cap H = \varnothing$. Тогда существует элемент $h \in H$ такой, что $h \in I$ и $h \notin I$. Следовательно, не все точки подмножества H, содержащиеся в интервале I, содержатся в интервале J. Противоречие с выполнением импликации $I \to J$.

Достаточность. Пусть

$$(I \backslash J) \cap H = \varnothing. \tag{4}$$

Отметим, что в силу изотонности оператора пересечения и включения $I \cap J \subseteq J$, выполняется включение

$$(I \cap J) \cap H \subseteq J \cap H. \tag{5}$$

Интервал I можно представить в виде $I = (I \backslash J) \cup (I \cap J)$.

Тогда $I \cap H = [(I \backslash J) \cup (I \cap J)] \cap H = [(I \backslash J) \cap H] \cup [(I \cap J) \cap H] = | (4) | = (I \cap J) \cap H \subseteq H | (5) | \subseteq I \cap H$.

Таким образом, $I \backslash H \subseteq J \cap H$. Следовательно, выполняется импликация $I \to J$. $\square$

Замечание. Из включений (2) и (5) следует выполнение равенства

$$I \cap H = (I \cap J) \cap H. \tag{6}$$

НЕКОТОРЫЕ СВОЙСТВА ИНТЕРВАЛЬНЫХ ИМПЛИКАТИВНЫХ ЗАВИСИМОСТЕЙ .

Пусть $I, J \in (B^n)$ - интервалы булевого пространства B^n, $H \subseteq B^n$.

Свойство 1. Если $I \subseteq J$, то импликация $I \to J$ выполняется.

Включение $I \subseteq J$ определяет тривиальный случай выполнения импликации $I \to J$, так как в этом случае, в силу изотонности операции пересечения, для любого подмножества H выполняется включение $I \cap H \subseteq J \subseteq H$.

Свойство 2. Если $I \cap H = \varnothing$, то для любого интервала $J \in (B^n)$ выполняется импликация $I \to J$.

Действительно,если $I \cap H = \varnothing$, то для любого интервала $J \in (B^n)$ выполняется включение $I \cap H = \varnothing \subseteq J \cap H$ и, следовательно, выполняется импликация $I \to J$.

Свойство 3. Если $I \cap H \neq \varnothing$ и $I \cap H = \varnothing$, то импликация $I \to J$ не выполняется.

Если $I \cap H \neq \varnothing$, то $(I \cap J) \cap H = \varnothing$. Следовательно, с учетом того, что $I \cap H \neq \varnothing$ получается неравенство $I \cap H \neq (I \cap J) \cap H$. Это противоречит равенству (6).

Свойство 4. Если выполняется импликация $I \to J$, то выполняется импликация $I \to (I \cap J)$.

Если выполняется импликация $I \to J$. Равенство (6) позволяет преобразовать это включение к виду $(I \cap J) \cap H \subseteq J \cap H$. Следовательно, $I \to (I \cap J)$.

Свойство 5. Если выполняется импликация $I \to J$, то для каждого интервала $K \in I(B^n)$, удовлетворяющего условию $I \cap K = I \cap J$, выполняется импликация $I \to K$.

Действительно, если выполняется импликация $I \to J$, то выполняется импликация $I \to (I \cap J)$, а так как $I \cap K = I \cap J$, то выполняется импликация $I \to (I \cap K)$. Поэтому, $I \cap H \subseteq (I \cap K) \cap H \subseteq K \cap H$, откуда следует выполнение импликации $I \to K$.

Перечисленные свойства импликативных зависимостей позволяют ограничиться рассмотрением только таких пар интервалов пространства B^n, для которых выполняется строгое включение $J \subset I$

ОЦЕНКА НЕСЛУЧАЙНОСТИ ИНТЕРВАЛЬНЫХ ИМПЛИКАТИВНЫХ ЗАВИСИМОСТЕЙ .

Способ оценки закономерности как неслучайности, предложенный А. Д. Закревским [1], позволяет, используя случайную независимую выборку, оценить размеры областей вида $I \backslash J$, которые можно практически достоверно считать свободными. Такая оценка выражается через вероятность того, что область оказалась свободной случайно. Если эта вероятность мала, но, не смотря на это, данные выборки порождают такую свободную область, то тогда велика вероятность того, что такая область свободна неслучайно (закономерно).

Пусть $V \subset G$ - случайная независимая выборка объема m из генеральной совокупности G.

$H \subseteq\in B^n$ - проекция выборки V в пространство B^n.

$I, J \in I(B^n)$ - фиксированные интервалы пространства B^n, для которых выполняются условия:

$$rang(I) = r \tag{7}$$

$$rang(J) = r + k, \tag{8}$$

где к ≥ 1.

Требование к ≥ 1 обеспечивает строгое включение $I \subset J$.

$I_{r,r+k}(B^n)$ - обозначение семейства всех пар (I, J) интервалов I и J, удовлетворяющих условиям (7) - (8) соответственно. Мощность семейства $I_{r,r+k}(B^n)$:

$$\mid I_{r,r+k}(B^n) \mid = C_n^r 2^r \cdot C_{n-r}^k 2^k. \tag{9}$$

Здесь $C_n^r 2^r$ - количество интервалов ранга r в пространстве B^n. $C_{n-r}^k 2^k$ - количество интервалов ранга $r+k$ в интервале ранга r.

Пусть для фиксированной пары интервалов (I, J) из семейства $I_{r,r+k}(B^n)$ на подмножестве H пространства B^n выполняется импликация $I \to J$. Определим вероятность случайного возникновения такой импликации. Для этого потребуется определить вероятность появления свободной области $I \backslash J$ на случайной независимой выборке объема m.

Фиксированные интервалы I и J, удовлетворяющие условиям (7) - (8), разбивают пространство B^n на три непересекающиеся области:

$$\overline{I} = B^n \backslash I, \tag{10}$$

$$J, \tag{11}$$

$$I \backslash J, \tag{12}$$

Предположим, что на множестве точек пространства B^n существует равномерное распределение вероятности. В этом случае, в силу ограничений (7) - (8):

$$P(I) = 2^{-r}, \tag{13}$$

$$P(\overline{I}) = 1 - 2^{-r}, \tag{14}$$

$$P(J) = 2^{-(r+k)}. \tag{15}$$

Событие «На случайной независимой выборке объема m из равномерного распределения μ точек выборки попадут в область J, l точек попадут в область $I \backslash J, m - (\mu + l)$ точек - в область $\overline{I}$» обозначим A. Определим вероятность такого события.

$$P(A) = C_m^{\mu} C_{m-\mu}^{l} P(J)^{\mu} P(I \backslash J)^{l} P(\overline{I})^{m-(\mu+l)}. \tag{16}$$

Область $I \backslash J$ должна остаться свободной от точек выборки. Поэтому полагаем $l = 0$. В этом случае равенство (16) примет вид:

$$P(A) = C_m^{\mu} C_{m-\mu}^{l} P(J)^{\mu} P(\overline{I})^{m-\mu}. \tag{17}$$

Или, с учетом равенств (14)-(15)

$$P(A) = C_m^{\mu} 2^{(l-\mu)(r+k)} (1 - 2^{-r})^{m-\mu}. \tag{18}$$

Оценим вероятность события «На случайной независимой выборке объема m существует хотя бы одна пара интервалов из семейства $I_{r,r+k}(B^n)$, для которой область $I\backslash J$ не содержит точек выборки (является свободной) и в интервале J содержится μ точек выборки». Обозначим это событие $\widetilde{A}(m,n,r,k,\mu)$. Такое событие представляет собой объединение всех событий A, порождаемых каждой парой интервалов из семейства $I_{r,r+k}(B^n)$:

$$\widetilde{A}(m,n,r,k,\mu) = \bigcup_{(I,J)\in I_{r,r+k}(B^n)} A. \tag{19}$$

Поэтому, с учетом количества элементов семейства $I_{r,r+k}(B^n)$ (равенство (9)),

$$P(\widetilde{A}(m,n,r,k,\mu)) \le \sum_{(I,J)\in I_{r,r+k}(B^n)} P(A) = W(m,n,r,k,\mu), \tag{20}$$

где

$$W(m,n,r,k,\mu) = C_n^r C_{n-r}^k C_m^\mu 2^{(l-\mu)(r+k)}(1-2^{-r})^{m-\mu}. \tag{21}$$

Ниже приведена таблица значений величины $W(m,n,r,k,\mu)$ при $n=100, m=200, \mu=10$ и некоторых значениях r и k.

$r\backslash k$	1	2	3	4	5	6	7	8
0	$00E+00$	00E + 00	00E + 00	00E + 00	00E + 00	00E + 00	00E + 00	00E + 00
1	$5,40E-43$	5,17E − 44	3,27E − 45	1,53E − 46	5,68E − 48	1,74E − 49	4,51E − 51	1,01E − 52
2	1,48E − 10	1,40E − 11	8,78E − 13	4,07E − 14	1,49E − 16	4,52E − 17	1,16E − 18	2,58E − 20
3	4,91E + 01	4,06E + 00	2,85E − 01	1,31E − 12	4,75E − 04	1,42E − 05	3,61E − 07	7,39E − 09
4	1,14E + 06	1,05E + 05	$6,45E+03$	2,93E + 02	1,05E + 01	3,12E − 01	7,82E − 03	1,70E − 04
5	2,14E + 07	1,96E + 06	1,19E + 05	5,34E + 03	1,90E + 02	5,56E + 00	1,38E − 01	2,97E − 03
6	1,37E + 07	1,24E + 06	7,44E + 04	3,31E + 03	1,16E + 02	3,37E + 00	8,27E − 02	1,16E − 03
7	1,59E + 06	1,43E + 05	8,49E + 03	3,37E + 02	1,30E + 01	3,72E − 01	9,02E − 03	1,89E − 04
8	7,56E + 04	6,72E + 03	3,94E + 02	1,71E + 01	5,88E − 01	1,66E − 02	3,99E − 04	8,29E − 06
9	2,17E + 03	1,90E + 02	1,1OE + 01	4,47E − 01	1,61E − 02	4,61E − 04	1,07E − 05	2,19E − 07
10	5,58E + 01	3,98E + 00	2,28E − 01	9,70E − 03	3,26E − 04	9,01E − 06	2,11E − 07	4,28E − 09
11	7,95E − 01	6,83E − 02	3,87E − 03	1,62E − 04	5,39E − 06	1,47E − 07	3,42E − 09	6,84E − 11
12	1,19E − 02	1,01E − 03	6,67E − 05	2,35E − 06	7,72E − 08	2,09E − 09	4,78E − 11	9,44E − 13
13	1,60E − 04	1,34E − 05	7,41E − 07	3,04E − 08	9,86E − 10	2,63E − 11	5,95E − 13	1,16E − 14
14	1,94E − 08	1,61E − 07	8,70E − O9	3,56E − 10	1,14E − И	3,01E − 13	6,72E − 15	1,30E − 16

Выражение (21) для величины W и приведенная таблица позволяют сделать следующие выводы.

1. При фиксированных значениях параметров r и μ значение величины W разбивают множество значений параметра k на два интервала: $[1,k_0]$ и $[k_0+1, n-r]$. При некоторых наборах значений параметров m,n,r,μ один из этих интервалов может отсутствовать.

В интервале $[1,k_0]$ величина W принимает значения больше единицы. Здесь оценка (21) не работает. В интервале $[k_0+1, n-r]$ величина W принимает значения меньше единицы и может рассматриваться как оценка вероятности события $\widetilde{A}(m,n,r,k,\mu)$.

Значение величины W уменьшается с ростом значения параметра k. Увеличение значения k означает увеличение ранга заключения импликации. Таким образом, с увеличением ранга заключения, вероятность случайного возникновения импликативной зависимости уменьшается, а, следовательно, увеличивается вероятность того, что эта импликация возникла неслучайно.

Такому поведению величины W может быть предложено следующее объяснение.

Пусть $l_1 l_2 \dots l_r$ - элементарная конъюнкция, соответствующая интервалу I, $l_1 l_2 \dots l_r m_1 m_2 \dots m_k$ - элементарная конъюнкция, соответствующая интервалу J.

Тогда описание свободной области $I \backslash J$ можно представить в виде

$$\bigvee_{j=1}^{k} l_1 l_2 \dots l_r \overline{m}_j \tag{22}$$

то есть свободная область $I \backslash J$ представляет собой объединение k свободных интервалов ранга $r+1$. Эти интервалы отделяют (изолируют) интервал j, содержащий μ точек выборки, по k направлениям от части (а может быть и от всех) остальных $m-\mu$ точек выборки. Чем больше значение k, тем по большему числу направлений (тем «лучше») изолирован интервал J, тем меньше вероятность того, что это произошло случайно.

2. При фиксированных значениях параметров k и μ значения величины W разбивают множество значений параметра r на три интервала: $[0, r_0], [r_0+1, r_1-1]$ и $[r_1, n-k]$. При некоторых наборах значений параметров m, n, k, μ один из этих интервалов может отсутствовать.

В интервале $[r_0+1, r_1-1]$ величина W принимает значения больше единицы. Здесь оценка (21) не работает. В интервалах $[0, r_0]$ и $[r_1, n-k]$ величина W принимает значения меньше единицы и может рассматриваться как оценка вероятности события $\widetilde{A}(m, n, r, k, \mu)$.

В интервале $[0, r_0]$ значение W возрастает с увеличением значения параметра r. Уменьшение значения параметра r означает уменьшение рангов посылки и заключения импликации. Таким образом, в интервале $[0, r_0]$ с уменьшением рангов посылки и заключения вероятность случайного возникновения импликативной зависимости уменьшается, а, следовательно, увеличивается вероятность того, что эта импликация возникла неслучайно. Такому поведению величины W может быть предложено следующее объяснение.

Пространственная мера интервалов, образующих свободную область, тем больше, чем меньше их ранг. Поэтому, чем меньше значение рангов этих интервалов, тем больше пространственная мера свободной области, отделяющей точки выборки, попавшие в интервал J от части (или всех) остальных точек выборки, тем меньше вероятность того, что это произошло случайно.

Следовательно, интервал $[0, r_0]$ определяет такую область изменения параметра r, в которой причиной возникновения импликативных зависимостей является достаточно большая пространственная мера свободной области, достаточно надежно отделяющей точки выборки, попавшие в интервал J от части (или всех) остальных точек выборки.

В интервале $[r_1, n-k]$ значение W убывает с увеличением значения параметра r. Увеличение значения параметра r означает увеличение рангов посылки и заключения импликации. Таким образом, в интервале $[r_1, n-k]$ с увеличением рангов посылки и заключения вероятность случайного возникновения импликативной зависимости уменьшается, а, следовательно, увеличивается вероятность того, что эта импликация возникла неслучайно. Такому поведению величины W может быть предложено следующее объяснение.

Пространственные меры интервалов I и J при увеличении их рангов становятся относительно небольшими. Вероятность того, что в такой относительно небольшой интервал J случайно попадет достаточно большая доля точек выборки, невелика. К этому добавляется еще то, что интервал J, содержащий μ точек выборки, отделен (изолирован) по k направлениям от части (а может быть и от всех) остальных $m-\mu$ точек выборки. Вероятность случайного появления такого события становится небольшой.

Следовательно, интервал $[r_1, n-k]$ определяет такую область изменения параметра r, в которой причиной возникновения импликативных зависимостей является то, что относительно небольшой интервал содержит достаточно большую долю точек выборки, и этот интервал хотя бы частично отделен от остальных точек выборки.

Таким образом, интервалы $[0, r_0]$ и $[r_1, n-k]$ соответствуют разным причинам возникновения импликативных зависимостей. Интервал $[0, r_0]$ характеризует области пространства $\in B^n$, которые «могут быть неплотно заселенные, но в достаточной степени хотя бы частично отделенные достаточно «большими» свободными областями». Интервал характеризует области, которые «небольшие, но плотно заселенные и частично отделенные свободными областями».

Интервал $[r_1, n-k]$ разделяет эти два интервала. Он соответствует случаю «неплотно заселенные и недостаточно отделенные свободными областями» области пространства.

При больших фиксированных значениях параметра k разделяющий интервал $[r_0+1, r_1-1]$ становится меньше, а при достаточно больших значениях k - вообще отсутствует. Однако разделение на два интервала, определяющих разные причины возникновения импликативной зависимости, сохраняется. На одном из этих интервалов величина W возрастает, на другом - убывает.

Основной результат статьи - введено понятие интервальной импликативной зависимости, и получена статистическая оценка неслучайности возникновения такой зависимости в булевых таблицах эмпирических данных.

Список литературы

1. Закревский А. Д. Логика распознавания. Мн.; Наука и техника, 1988. - 118с.
2. Ganter, B.. Wille, R. Formal Concept Analysis - Mathematical Foundations. Springer - Verlag, Berlin., 1999.
3. Mohammed J.Zaki Mining non-redundant association rules. Data Min. Knowl. Discov., 9(3):223-248, 2004. «Таврический вестник информатики и математики», Ml 2006