Обработка данных для индуктивного вывода на основе нестрогой вероятности

Л. В. Аршинский, В. С. Лебедев

Иркутский государственный университет путей сообщения, Иркутск, Россия

Аннотация. На основе методов индуктивной логики рассматривается подход к выявлению импликативных взаимосвязей вида «Если A, то b» в больших данных в условиях их низкой достоверности и противоречивости. Для работы с данными используются логики с векторной семантикой в форме VTF-логик. Наличие или отсутствие явлений в таблицах их совместной встречаемости формализуется векторами истинности с компонентами v+ и v–, где v+ мера истинности утверждения о наличии явления, v– - мера его ложности. На основе статистической индукции вычисляется показатель обоснованности причинно-следственной связи как усредненное значение векторов истинности соответствующих нестрогих высказываний. Получившееся значение трактуется как нестрогая вероятность связи, которая выступает векторным показателем ее обоснованности. Обсуждается применимость подхода для обработки качественных и количественных данных, а также данных, содержащих артефакты.

Ключевые слова: большие данные, интеллектуальный анализ данных, индуктивный вывод, нестрогая вероятность, логики с векторной семантикой.

DOI 10.14357/20718632240201 **EDN** HUCOJV

Введение

В настоящее время рост мощностей вычислительной техники и технологий передачи информации по проводным и беспроводным каналам связи приводит к накоплению огромных объемов информации – больших данных. Большие данные представляют собой структурированные или неструктурированные массивы данных большого объема и характеризуются набором из семи V: volume, velocity, variety, veracity, variability, visualization, value (объем, скорость, разнообразие, достоверность, изменчивость, визуализация, ценность) [1]. Говорят также о «тройках», «четверках», «пятерках V» [2; 3]. Целью работы с такими данными является выявление в них взаимосвязей, знание которых может быть положено в основу тех или иных управленческих решений [4; 5]. Часто в этом качестве выступают скрытые закономерности. Их извлечение, например, средствами Data Mining – интеллектуального анализа данных (ИАД), – одно из основных направлений анализа больших данных (АБД) [6; 7].

Особенно стремительный рост объема данных начался с нулевых годов, что повлияло на стоимость их хранения и методы работы с ними [8]. Возникла необходимость в новых инструментах, позволяющих более эффективно анализировать растущие объемы информации. Для таких данных постоянно разрабатываются новые способы обработки в расчете на улучшение соответствующих методов и технологий [9]. Область применения больших данных продолжает расширяться. На текущий момент они нашли применение в бизнесе, банковской сфере, ретейле, маркетинге, госструктурах, логистике,

машиностроении, медицине. Пример их использования для принятия управленческих решений в государственном управлении описан в [10]. Вообще, в литературе уже представлено довольно много случаев, когда в результате подобного анализа обнаруживали зависимости, существенные для ведения успешной деятельности в той или иной предметной области [7].

Считается, что АБД включает в себя два основных этапа [11]:

1) предобработку, где данные очищаются и приводятся к виду, позволяющему применять тот или иной метод или группу методов;

2) собственно анализ, где из данных извлекается информация, которая является целью анализа; например, сведения о скрытых закономерностях.

Обычно в ходе предобработки из массива данных удаляются малодостоверные и противоречивые фрагменты, заполняются пропуски, выявляются сомнительные аномалии и иные артефакты [12]. При этом отказываются от анализа, если объем некачественной информации превышает некоторый порог, к примеру, 20% от общего количества [13]. Такой прием увеличивает доверие к результатам анализа, однако неудобен тем, что пока порог не преодолен, все данные, даже исправленные, предполагаются достоверными, а как только он преодолевается, все зачисляются в сомнительные. Своего рода «релейное» управление процессом.

Представляется, что «релейности» можно было бы избежать, или, как минимум, ослабить ее влияние, если вместе с самими данными учитывать степень доверия к ним, задав ее некоторым числом, скажем, из интервала [0,1]. Здесь 0 означает, что доверия к элементу данных (скажем, записи в базе данных или соответствующему полю в ней) нет, а 1 – безусловное доверие. В этом случае извлеченные взаимосвязи сопровождаются показателями, которые можно рассматривать как показатели обоснованности того или иного результата. Изменение числа и степени влияния артефактов в этом случае просто меняет обоснованность заключения, подсказывая насколько подкреплено получающееся знание. Учитывая, что часть данных может выступать в качестве аргументов в пользу соответствующей гипотезы, а часть опровергать ее, одновременно можно учитывать и степень ее противоречивости, а также степень определенности с точки зрения доверия к соответствующим фрагментам данных, точнее – их источникам.

Проблема обработки неполной и противоречивой информации при анализе данных, в том числе с помощью индуктивных методов рассматривается не впервые. Достаточно упомянуть целый класс работ по ДСМ-методу В.К. Финна как самого автора метода, так и его последователей [14-18]. Вполне основательно это направление освещено в [18]. Особенностью метода является рассмотрение «позитивных» (+), «негативных» (–), «фактически противоречивых» (0) и «фактически неопределенных» (τ) примеров, влияющих на принятие/непринятие гипотез. В известном смысле рассуждения строятся в четырехзначной семантике Данна и Белнапа [19-22]. Однако при этом примеры зачисляются в строго положительные, строго отрицательные, неопределенные, либо полностью противоречивые множества, что, как представляется, огрубляет ситуацию. Более аккуратно такие особенности данных можно учесть, если опираться на векторную семантику в форме семантики V^{TF}, впервые описанную в [14] и ряде более поздних работ.

1. Описание подхода

Одним из традиционных подходов к анализу данных с целью извлечения закономерностей является индуктивный вывод. Он включает ряд техник (правил вывода), одними из основных среди которых выступают соединенный метод сходства и различия (СМСР) и метод единственного сходства (МЕС) [24]. Фактически, они основаны на анализе таблиц совместной встречаемости, подобных Табл. 1.

В обоих методах a_i и b – это качественные показатели (явления), зарегистрированные в ходе экспериментов и которые имеются или отсутствуют (значения a_{ki} и b_k – это 1 либо 0). Общее число проведенных опытов K. Требуется обосновать зависимость вида:

$$A(a_{i1},...,a_{ij}) \to b \ (\text{«Если } A(a_{i1},...,a_{ij}) \text{, то } b\text{»}), \quad (1)$$

Табл. 1. Совместная встречаемость явлений

	a_1	a_2	…	a_n	b
1	a_{11}	a_{21}	…	a_{n1}	b_1
2	a_{12}	a_{22}	…	a_{n2}	b_2
3	a_{13}	a_{23}	…	a_{n3}	b_3
…	…	…	…	…	…
K	a_{1K}	a_{2K}	…	a_{nK}	b_K

где $A(a_{i1},...,a_{ij})$ булево выражение, состоящее из некоторого подмножества множества показателей $\{a_{1,...,}a_n\}$, объединенных связками конъюнкции (&), дизъюнкции ($\vee$), отрицания ($\neg$) и, возможно, содержащее скобки. В простейшем случае это $a_i \rightarrow b$.

Согласно СМСР, для обоснования нужно удостовериться, что в каждой строке таблицы истинно высказывание:

$$A(a_{i1},...a_{ij}) \,\&\, b \vee \neg A(a_{i1},...a_{ij}) \,\&\, \neg b\,. \quad (2)$$

Остальные элементы таблицы могут принимать произвольные значения (строго говоря, в классическом СМСР строки, в которых $A(a_{i1},...,a_{ij})$ истинно, должны отличаться от строк, где $A(a_{i1},...,a_{ij})$ ложно только наличием отрицания при $A(\ldots)$, однако в реальных, не искусственных базах данных это не всегда выполняется в связи с чем СМСР рассматриваем в указанной форме).

Если (2) истинно для всех строк, зависимость (1) полагаем обоснованной и правдоподобной. Если же (2) окажется истинным только для части строк, тогда отношение их числа, где (2) истинно, к общему количеству K можно рассматривать как меру обоснованности (2). В индуктивной логике такой прием связывают с вероятностью и статистической индукцией [24-27].

Для МЕС выражение (2) принимает более простой вид:

$$A(a_{i1},...a_{ij}) \,\&\, b\,. \quad (3)$$

Описанный прием успешно работает, когда все a_i и b известны с достоверностью. В этом случае каждый показатель принимает значение 0 или 1 и установить истинность (2) и (3) несложно. Однако, когда полной уверенности нет, когда источники данных малодостоверны или противоречивы, требуется искать иные подходы. В частности, такое может возникнуть в ходе устранения артефактов, что характерно для АБД.

В [28] для подобных случаев описан прием, основанный на понятии *нестрогой вероятности* [29]. Под нестрогой вероятностью случайного события A понимается вектор:

$$P(A) = \langle P^+(A); P^-(A)\rangle = \\ = \langle \sum_{\omega\in\Omega} F^+(\omega, A)\,p(\omega); \sum_{\omega\in\Omega} F^-(\omega, A)\,p(\omega)\rangle\,. \quad (4)$$

Здесь $F^+(\omega,A)\in[0,1]$ – позитивный компонент вектора истинности утверждения:

$$F(\omega,A) = \text{«Элементарное событие } \omega \text{ благоприятно с точки зрения события } A\text{»}, \quad (5)$$

показывающий насколько (5) есть истинно, а $F^-(\omega,A)\in[0,1]$ – негативный компонент этого вектора, показывающий насколько оно ложно; $p(\omega)$ – вероятность ω в обычном смысле; Ω – полная группа элементарных событий. Истинность и ложность определяется подтверждающими и опровергающими свидетельствами. Если их рассматривать как независимые, получаем вектор истинности

$$||F(\omega,A)|| = \langle F^+(\omega,A); F^-(\omega,A)\rangle,$$

где $0 \le F^+(\omega,A)+F^-(\omega,A) \le 2$.

Представление о нестрогой вероятности имеет смысл, когда нет твердой уверенности в благоприятности/неблагоприятности ω для A, но есть (независимые) доводы «за» и «против» с разной степенью доверия к ним. Для строгих значений вектора $||F(\omega,A)||$, равных $\langle 1;0\rangle$ (строгая истина) или $\langle 0;1\rangle$ (строгая ложь), (4) превращается в привычную вероятность, где $P^+(A)$ – вероятность A, а $P^-(A)$ – вероятность противоположного события.

Следует обратить внимание, что $P^+(A)$ и $P^-(A)$ также не зависят друг от друга.

С точки зрения обсуждаемого вопроса, событие ω – это запись в Табл. 1. Все записи образуют множество Ω. Вероятности $p(\omega)$ считаем одинаковыми и равными $1/K$.

В случае недостатка уверенности в том или ином показателе, или получении как подтверждающих, так и опровергающих свидетельств (противоречие), Табл. 1 принимает форму [20] (Табл.2). Здесь $a_{ik}^+, a_{ik}^- \in [0,1]$. В соответствии с (2) значение вектор-функции $\|F(\omega,A)\|$ для каждой строки k вычисляется как:

$$\| A(a_{i_1 k},...,a_{i_j k}) \,\&\, b_k \vee \neg A(a_{i_1 k},...,a_{i_j k}) \,\&\, \neg b_k \| . \quad (6)$$

Для (3) это выглядит как:

$$\| A(a_{i_1 k},...,a_{i_j k}) \,\&\, b_k \| . \quad (7)$$

Истинность антецедента $A(a_{i1k},...,a_{i\,jk})$ рассчитывается согласно:

$$\|v\&u\| = \langle v^+ \bullet u^+;\ v^- \oplus u^- \rangle;$$
$$\|v \vee u\| = \langle v^+ \oplus u^+;\ v^- \bullet u^- \rangle;$$
$$\|\neg v\| = \langle v^-; v^+ \rangle,$$

– первые формы конъюнкции, дизъюнкции и отрицания [23]. Приоритет связок обычный.

Вектор нестрогой вероятности в случае СМСР:

$$P(A) = \langle \frac{1}{K}\sum_{k=1}^{K}[A_k^+ \bullet b_k^+ \oplus A_k^- \bullet b_k^-];\ \frac{1}{K}\sum_{k=1}^{K}[(A_k^- \oplus b_k^-) \bullet (A_k^+ \oplus b_k^+)] \rangle \quad (8)$$

характеризует обоснованность гипотезы (1) с учетом доверия к полученной информации, если она недостаточно убедительна и/или противоречива. Для МЕС выражение выглядит проще:

$$P(A) = \langle \frac{1}{K}\sum_{k=1}^{K}[A_k^+ \bullet b_k^+]; \frac{1}{K}\sum_{k=1}^{K}[(A_k^- \oplus b_k^-)] \rangle \quad (9)$$

Символы $x \bullet y$ и $x \oplus y$ – треугольная норма $t(x,y)$ и ко-норма $s(x,y)$ [30; 31] в инфиксной записи с дополнительно введенной аксиомой:

$$(1-x) \bullet (1-y) + x \oplus y = 1; \quad (10)$$

приоритет: $\bullet$, $\oplus$, $\pm$. Треугольные нормы, дополненные аксиомой связи (11) в [23] названы композиционным умножением и сложением (по аналогии с обычными умножением и сложением, но на отрезке [0,1]). Типичными примерами таких норм являются:

$$x \bullet y = \min(x,y);\ x \oplus y = \max(x,y). \quad (11)$$
$$x \bullet y = xy;\ x \oplus y = x + y - xy; \quad (12)$$

Значение векторов истинности $\langle a_{ik}^+; a_{ik}^- \rangle$ получаются из следующих соображений:

1. *По степени доверия к источникам*. Пусть в первой строке наличие показателя a_1 совместно утверждается и отрицается двумя разными источниками. Если первому мы доверяем в степени 0.9, а другому 0.7, вектор $\langle a_{11}^+; a_{11}^- \rangle$ принимает значение $\langle 0.9;0.7 \rangle$. Формально, это объединение двух векторов $\langle 0.9;0 \rangle$ и $\langle 0;0.7 \rangle$ по правилу:

$$\|v\| = \langle v_1^+ \oplus v_2^+; v_1^- \oplus v_2^- \rangle \quad (13)$$

(или $\langle v_1^+ \oplus \ldots \oplus v_q^+; v_1^- \oplus \ldots \oplus v_q^- \rangle$, если источников больше). Подобное возможно, например, при получении различающихся результатов лабораторных исследований, когда доверие к результатам разное, а также при наличии артефактов-противоречий.

Если неполную уверенность в надежности источников трактовать как частичную уверенность в противоположном исходе, объединяются вектора $\langle 0.9;0.1 \rangle$ и $\langle 0.3;0.7 \rangle$. Для s-нормы в виде $x \oplus y = \max(x,y)$ это вновь $\langle 0.9;0.7 \rangle$. Если степень доверия к источнику не определена (информация отсутствует), вектор можно задать

Табл. 2. Совместная встречаемость явлений в векторном представлении

	a_1	a_2	…	a_n	B
1	$\langle a_{11}^+; a_{11}^- \rangle$	$\langle a_{21}^+; a_{21}^- \rangle$	…	$\langle a_{n1}^+; a_{n1}^- \rangle$	$\langle b_1^+; b_1^- \rangle$
2	$\langle a_{12}^+; a_{12}^- \rangle$	$\langle a_{22}^+; a_{22}^- \rangle$	…	$\langle a_{n2}^+; a_{n2}^- \rangle$	$\langle b_2^+; b_2^- \rangle$
3	$\langle a_{13}^+; a_{13}^- \rangle$	$\langle a_{23}^+; a_{23}^- \rangle$	…	$\langle a_{n3}^+; a_{n3}^- \rangle$	$\langle b_3^+; b_3^- \rangle$
…	…	…	…	…	…
K	$\langle a_{1K}^+; a_{1K}^- \rangle$	$\langle a_{2K}^+; a_{2K}^- \rangle$	…	$\langle a_{nK}^+; a_{nK}^- \rangle$	$\langle b_K^+; b_K^- \rangle$

как ⟨0.5;0.5⟩. В последних двух случаях справедливо соотношение:

$$a_{ik}^{+}+a_{ik}^{-}=1,$$

– частный случай векторного представления истинности.

2. *По статистическим соображениям*. Этот подход может применяться при замещении артефактов некоторыми усредненными показателями. В качестве примера возьмем пропущенное поле «пол» в БД сотрудников/клиентов компании. Если нет других свидетельств, вычисляем соотношение мужчин и женщин, формируя, к примеру, вектор ⟨0.52;0.48⟩.

3. *Объединение обоих подходов*. В этом случае вектор формируется из статистических соображений, но далее каждый компонент умножается на степень доверия к искусственно введенному показателю. К примеру, 0.5 (тогда предыдущий вектор примет значение ⟨0.26;0.24⟩). Этот простой эвристический прием позволит управлять влиянием артефактов на итоговый результат, причем он предполагает именно векторное представление истинности.

Возможны и другие стратегии, но все они дают общий результат: малодостоверные, сомнительные данные снижают обоснованность гипотезы. Причем она уменьшается постепенно с ростом объема ненадежных данных. Внешне это выражается в изменении показателей:

достоверности (обоснованности)
$$\mu_{д}(A) = P(A^{+})–P(A^{-});$$
определенности
$$\mu_{о}(A) = P(A^{+})\oplus P(A^{-});$$
противоречивости
$$\mu_{п}(A) = P(A^{+})\bullet P(A^{-});$$

и некоторых других [29], что позволит более гибко управлять анализом. Выбор из нескольких альтернативных гипотез может выполняться на основе лексикографического порядка $\{\mu_{д},\mu_{о},1-\mu_{п}\}$.

Особенностью подхода является то, что в ячейках Табл. 1 и производной от нее Табл. 2 представлены качественные значения 0/1, тогда как на практике соответствующие таблицы могут содержать и количественные показатели. Попробуем учесть это обстоятельство.

2. Обработка количественных данных

Для перехода от числовых данных к качественным воспользуемся следующим приемом. Разделим весь диапазон числовых значений на множество непересекающихся поддиапазонов так, что любое возможное значение попадет в один из них. Количество диапазонов определяется гипотезами, которые ставит исследователь. Например, можно рассматривать гипотезы о связи фактора b с превышением/непревышением соответствующим показателем некоторого порога, попадании числа в допустимый/недопустимый диапазон значений и т.п. Возможны другие варианты.

Разберем описанную выше схему подробнее на случаях с двумя и тремя гипотезами. При использовании двух гипотез берется пороговое значение T и все значения данных делятся на два диапазона, например, $(-\infty;T]$ и $(T,+\infty)$. Попадание/непопадание в диапазон – это наличие или отсутствие единственного качества. Соответствующее утверждение, принимающее векторное значение истинности, это:

h_1=«Числовое значение не превышает порог T»;
либо
h_2=«Числовое значение превышает порог T».

Очевидно, что h_2 это $\neg h_1$ и наоборот. Переход к нестрогой вероятности для такого случая, фактически, обсужден выше.

Интереснее случай трех подинтервалов: ниже нормы/норма/выше нормы. Здесь выбираются два пороговых значения T_1 и T_2, с помощью которых весь числовой диапазон делится, к примеру, так: $(-\infty,T_1)$, $[T_1,T_2]$, $(T_2,+\infty)$. Это соответствует трем возможным качествам (Табл. 3):
h_1= «Числовое значение не превышает порог T_1»;
h_2= «Числовое значение находится в границах между T_1 и T_2»;
h_3= «Числовое значение превышает порог T_2».

Векторное представление Табл. 3 показано на Табл. 4.

Здесь учтен факт, что числовое значение может принадлежать только одному из интервалов.

Разберем работу с векторами истинности в случае неполной уверенности в данных.

Табл. 3. Попадание числа в один из трех диапазонов

Значение показателя	h_1	h_2	h_3	B
$v \le T_1$	1	0	0	0
$T_1 < v \le T_2$	0	1	0	1
$v > T_2$	0	0	1	0

Табл. 4. Векторное представление истинности для трех диапазонов

Значение показателя	h_1	h_2	h_3	B
$v \le T_1$	$\langle 1;0\rangle$	$\langle 0;1\rangle$	$\langle 0;1\rangle$	$\langle 0;1\rangle$
$T_1 < v \le T_2$	$\langle 0;1\rangle$	$\langle 1;0\rangle$	$\langle 0;1\rangle$	$\langle 1;0\rangle$
$v > T_2$	$\langle 0;1\rangle$	$\langle 0;1\rangle$	$\langle 1;0\rangle$	$\langle 0;1\rangle$

1. *По степени доверия к источникам*. Предположим, источник сообщает о принадлежности числа к интервалу $(-\infty, T_1)$ (качество h_1) и доверие к источнику имеет величину α. Это означает недоверие той же величины к качествам h_2 и h_3 (свидетельство силы α в пользу h_1 есть свидетельство той же силы против h_2 и h_3). Согласно этому получаем следующие значения истинности: $\|h_1\|=\langle \alpha;0\rangle$, $\|h_2\|=\langle 0;\alpha\rangle$, $\|h_3\|=\langle 0;\alpha\rangle$ [32]. Этот вариант иллюстрируется Табл. 5.

Для полноты следовало бы упомянуть случай, когда источник с доверием α сообщает о непринадлежности числа интервалу h_1: $\|h_1\|=\langle 0;\alpha\rangle$ (истинности h_2 и h_3 здесь принимают интервальные значения: $\|h_2\|=\|h_3\|=\langle [0,\alpha];0\rangle$ [26]), но он вряд ли интересен с прикладной точки зрения.

2. *По статистическим соображениям*. Здесь истинность определяется относительной частотой попадания в каждый из интервалов. Обозначая частоты, соответственно, α_1, α_2 и α_3, запишем: $\|h_1\|=\langle \alpha_1;\alpha_2+\alpha_3\rangle$, $\|h_2\|=\langle \alpha_2;\alpha_1+\alpha_3\rangle$, $\|h_3\|=\langle \alpha_3;\alpha_1+\alpha_2\rangle$; $\alpha_1+\alpha_2+\alpha_3=1$. Данный прием подходит при работе с артефактами.

3. *Объединение обоих подходов*. Может применяться также при замене артефактов искусственными, например, статистически определенными значениями. Как описано выше, формируемые векторы умножаются на степень доверия к искусственному показателю, что снижает влияние таких показателей на результат. При нулевом доверии влияние артефактов исключается вовсе.

В результате для каждой гипотезы получаются группы векторов для каждого набора данных и к ним можно применить описанную технику. Она пригодна как для СМСР, так и для МЕС. Проиллюстрируем это расчетом.

3. Примеры расчета

Расчет с помощью СМСР в форме (2) проведем на гипотетических данных, что позволяет показать особенности вычислений. Для иллюстрации МЕС воспользуемся реальным массивом данных по заболеванию диабетом, представленным в [33]. Различие в расчетах между СМСР и МЕС состоит в замене (2) на (3) и, соответственно, (6) и (9) на (7) и (10). Все остальное сохраняется. Выбор МЕС во втором примере обусловлен тем, что соответствующие данные содержат сведения о достаточно специфичной группе людей, что при СМСР может привести к некорректным результатам (хотя любая индукция в принципе требует осторожности).

Табл. 5. Векторное представление при ограниченном доверии к данным

Значение показателя	Доверие	h_1	h_2	h_3	b
$v \le T_1$	0.8	$\langle 0.8;0\rangle$	$\langle 0;0.8\rangle$	$\langle 0;0.8\rangle$	$\langle 0;1\rangle$
$T_1 < v \le T_2$	0.8	$\langle 0;0.8\rangle$	$\langle 0.8;0\rangle$	$\langle 0;0.8\rangle$	$\langle 0;1\rangle$
$v > T_2$	0.8	$\langle 0;0.8\rangle$	$\langle 0;0.8\rangle$	$\langle 0.8;0\rangle$	$\langle 1;0\rangle$

Часто при проведении медицинских анализов значение какого-либо показателя сравнивается с показателями здорового человека. Обычно такие значения оцениваются по порогу или заключаются в диапазон от нижней до верхней границы. В качестве примера возьмем уровень глюкозы в крови и свяжем его с наличием/отсутствием диабета 2 типа. Норма глюкозы у человека 3,3-5,5 ммоль/л [34]. Предположим, получены результаты из Табл. 6. Здесь значение 0 и 1 сразу заменены векторами ⟨0;1⟩ и ⟨1;0⟩ – строгая ложь и строгая истина, этим переводим скалярные значения истинности в векторное представление, необходимое для рассматриваемой техники.

В Табл. 6 доверие к показаниям полное. Рассмотрим случай, когда доверие к ним в силу каких-то обстоятельств меньше единицы. Это ситуация 1 из рассмотренных выше. Результат представлен в Табл. 7.

Исследуются гипотезы:

$H_1 = h_1 \rightarrow$диабет 2 типа;
$H_2 = h_2 \rightarrow$диабет 2 типа;
$H_3 = h_3 \rightarrow$диабет 2 типа.

Вычисляем истинности гипотез согласно (6) и нестрогую вероятность согласно (8) (Табл. 8). Треугольные нормы (композиционное умножение и сложение) выбираем как в (11), округление до тысячных.

Полученные значения нестрогих вероятностей:

$P(H_1)=\langle 0.188;0.438\rangle$;
$P(H_2)=\langle 0.3;0.325\rangle$;
$P(H_3)=\langle 0.6;0.025\rangle$.

Меры достоверности и определенности для гипотез принимают значения:

$\mu_д(H_1)= -0.25$, $\mu_о(H_1)=0.438$, $\mu_п(H_1)=0.188$;
$\mu_д(H_2)= -0.025$, $\mu_о(H_2)=0.325$, $\mu_п(H_2)=0.3$;
$\mu_д(H_3)=0.575$, $\mu_о(H_3)=0.6$, $\mu_п(H_3)=0.025$.

Используя лексикографический порядок $\{\mu_д,\mu_о,1-\mu_п\}$ останавливаемся на гипотезе H_3.

Рассмотрим вариант, когда часть строк, например первая и третья, сформированы из разных источников, которые предоставили взаимоисключающие данные. Объединяя свидетельства согласно (13), получаем Табл. 9. S-норму берем также по (11), это дает результат из Табл. 10.

Табл. 6. Взаимосвязь уровня глюкозы с диабетом в векторном представлении

	Уровень глюкозы	h_1	h_2	h_3	Диабет 2 типа
1	4.1	⟨0;1⟩	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩
2	2.9	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
3	6.5	⟨0;1⟩	⟨0;1⟩	⟨1;0⟩	⟨1;0⟩
4	3.1	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
5	5.3	⟨0;1⟩	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩
6	7.0	⟨0;1⟩	⟨0;1⟩	⟨1;0⟩	⟨1;0⟩
7	3.2	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
8	4.6	⟨0;1⟩	⟨1;0⟩	⟨0;1⟩	⟨1;0⟩

Табл. 7. Переход к векторному представлению при неполном доверии к данным

	Уровень глюкозы	Доверие	h_1	h_2	h_3	Диабет 2 типа
1	4.1	0.8	⟨0;0.8⟩	⟨0.8;0⟩	⟨0;0.8⟩	⟨0;1⟩
2	2.9	0.3	⟨0.3;0⟩	⟨0;0.3⟩	⟨0;0.3⟩	⟨0;1⟩
3	6.5	0.6	⟨0;0.6⟩	⟨0;0.6⟩	⟨0.6;0⟩	⟨1;0⟩
4	3.3	0.9	⟨0.9;0⟩	⟨0;0.9⟩	⟨0;0.9⟩	⟨0;1⟩
5	5.3	0.7	⟨0;0.7⟩	⟨0.7;0⟩	⟨0;0.7⟩	⟨0;1⟩
6	7.0	0.5	⟨0;0.5⟩	⟨0;0.5⟩	⟨0.5;0⟩	⟨1;0⟩
7	3.2	1.0	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
8	4.6	0.2	⟨0;0.2⟩	⟨0.2;0⟩	⟨0;0.2⟩	⟨1;0⟩

Табл. 8. Нестрогие вероятности гипотез H_1, H_2, H_3

	Уровень глюкозы	Доверие	H_1	H_2	H_3
1	4.1	0.8	⟨0.8;0⟩	⟨0;0.8⟩	⟨0.8;0⟩
2	2.9	0.3	⟨0;0.3⟩	⟨0.3;0⟩	⟨0.3;0⟩
3	6.5	0.6	⟨0;0.6⟩	⟨0;0.6⟩	⟨0.6;0⟩
4	3.3	0.9	⟨0;0.9⟩	⟨0.9;0⟩	⟨0.9;0⟩
5	5.3	0.7	⟨0.7;0⟩	⟨0;0.7⟩	⟨0.7;0⟩
6	7.0	0.5	⟨0;0.5⟩	⟨0;0.5⟩	⟨0.5;0⟩
7	3.2	1.0	⟨0;1⟩	⟨1;0⟩	⟨1;0⟩
8	4.6	0.2	⟨0;0.2⟩	⟨0.2;0⟩	⟨0;0.2⟩
Нестрогая вероятность:			⟨0.188;0.438⟩	⟨0.3;0.325⟩	⟨0.6;0.025⟩

Табл. 9. Векторное представление при взаимоисключающих свидетельствах

	Уровень глюкозы	Доверие	h_1	h_2	h_3	Диабет 2 типа
1	4.1; 7.2	0.8; 0.3	⟨0;0.8⊕0.3⟩	⟨0.8;0⊕0.3⟩	⟨0.3;0.8⟩	⟨0;1⟩
2	2.9	0.3	⟨0.3;0⟩	⟨0;0.3⟩	⟨0;0.3⟩	⟨0;1⟩
3	6.5; 3.1	0.6; 0.4	⟨0.4;0.6⟩	⟨0;0.6⊕0.4⟩	⟨0.6;0⊕0.4⟩	⟨1;0⟩
4	3.3	0.9	⟨0.9;0⟩	⟨0;0.9⟩	⟨0;0.9⟩	⟨0;1⟩
5	5.3	0.7	⟨0;0.7⟩	⟨0.7;0⟩	⟨0;0.7⟩	⟨0;1⟩
6	7.0	0.5	⟨0;0.5⟩	⟨0;0.5⟩	⟨0.5;0⟩	⟨1;0⟩
7	3.2	1.0	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
8	4.6	0.2	⟨0;0.2⟩	⟨0.2;0⟩	⟨0;0.2⟩	⟨1;0⟩

Табл. 10. Векторное представление при взаимоисключающих свидетельствах

	Уровень глюкозы	Доверие	h_1	h_2	h_3	Диабет 2 типа
1	4.1; 7.2	0.8; 0.3	⟨0;0.8⟩	⟨0.8; 0.3⟩	⟨0.3;0.8⟩	⟨0;1⟩
2	2.9	0.3	⟨0.3;0⟩	⟨0;0.3⟩	⟨0;0.3⟩	⟨0;1⟩
3	6.5; 3.1	0.6; 0.4	⟨0.4;0.6⟩	⟨0;0.6⟩	⟨0.6;0.4⟩	⟨1;0⟩
4	3.3	0.9	⟨0.9;0⟩	⟨0;0.9⟩	⟨0;0.9⟩	⟨0;1⟩
5	5.3	0.7	⟨0;0.7⟩	⟨0.7;0⟩	⟨0;0.7⟩	⟨0;1⟩
6	7.0	0.5	⟨0;0.5⟩	⟨0;0.5⟩	⟨0.5;0⟩	⟨1;0⟩
7	3.2	1.0	⟨1;0⟩	⟨0;1⟩	⟨0;1⟩	⟨0;1⟩
8	4.6	0.2	⟨0;0.2⟩	⟨0.2;0⟩	⟨0;0.2⟩	⟨1;0⟩

Соответственно, Табл. 8 принимает вид Табл. 11.

Меры достоверности и определенности для гипотез принимают значения:

$\mu_д(H_1)= -0.2$, $\mu_о(H_1)=0.438$, $\mu_п(H_1)=0.238$;
$\mu_д(H_2)=0.013$, $\mu_о(H_2)=0.338$, $\mu_п(H_2)=0.325$;
$\mu_д(H_3)=0.488$, $\mu_о(H_3)=0.6$, $\mu_п(H_3)=0.112$.

Видно, что повысились достоверности первой и второй гипотез и упала достоверность третьей, что вполне оправдано. Показатели определенности и противоречия поменялись тоже.

Для иллюстрации подхода на реальных данных используем таблицу из [33]. При этом интересно сравнить результат, полученный в предположении их достоверности (т.е. обычная статистическая индукция), и с учетом влияния артефактов. Достоверным выглядит столбец значения глюкозы (Glucose) в связи с диагностированным заболеванием. Гипотеза: «Высокое значение глюкозы→Диабет». Рассматривая для МЕС только строки, где представлены больные, на основе (10) получаем с точностью до третьего

Табл. 11. Нестрогие вероятности гипотез H_1, H_2, H_3 при взаимоисключающих свидетельствах

	Уровень глюкозы	Доверие	H_1	H_2	H_3
1	4.1	0.8	⟨0.8;0⟩	⟨0.3;0.8⟩	⟨0.8;0.3⟩
2	2.9	0.3	⟨0;0.3⟩	⟨0.3;0⟩	⟨0.3;0⟩
3	6.5	0.6	⟨0.4;0.6⟩	⟨0;0.6⟩	⟨0.6;0.4⟩
4	3.3	0.9	⟨0;0.9⟩	⟨0.9;0⟩	⟨0.9;0⟩
5	5.3	0.7	⟨0.7;0⟩	⟨0;0.7⟩	⟨0.7;0⟩
6	7.0	0.5	⟨0;0.5⟩	⟨0;0.5⟩	⟨0.5;0⟩
7	3.2	1.0	⟨0;1⟩	⟨1;0⟩	⟨1;0⟩
8	4.6	0.2	⟨0;0.2⟩	⟨0.2;0⟩	⟨0;0.2⟩
Нестрогая вероятность:			⟨0.238;0.438⟩	⟨0.338;0.325⟩	⟨0.6;0.112⟩

знака значение вектора обоснованности $P = \langle 0.996;0\rangle$; t- и s-нормы брались согласно (12). Статистическая индукция дает значение 1. Практически полное совпадение, что объясняется небольшим числом артефактов (нулевых значений) по этому показателю.

Иной результат дает взаимосвязь «Инсулин-Диабет». Полагаем, что артефакты здесь представлены нулевыми значениями столбца Insulin. Доля подобных артефактов составляет около 48,7%. Для работы с ними использовалась простая эвристика: нулевые значения заменялись средним по ненулевым показателям столбца с последующим выставлением для таких ячеек значения доверия 0.5. В результате получены векторные значения обоснованности гипотез (нестрогие вероятности) по МЕС:

P(Инсулин ниже нормы→Диабет) = ⟨0.007;0.735⟩;
P(Инсулин в норме→Диабет) = ⟨0.627;0.116⟩;
P(Инсулин выше нормы→Диабет) =⟨0.108;0.634⟩.

Если же в таблице вновь оставить только больных, а потом удалить строки с артефактами по инсулину, получаем статистику:

P(Инсулин ниже нормы→Диабет) = 0.007;
P(Инсулин в норме→Диабет) = 0.884;
P(Инсулин выше нормы→Диабет) = 0.108

(известный факт, что диабет может наблюдаться при нормальном уровне инсулина, если это диабет второго типа – наиболее массовый; о причинно-следственной связи здесь речь не идет). И в этом примере наблюдается корреляция между результатами на основе нестрогой вероятности и статистикой. Но нестрогие вероятности получены для почти пятидесятипроцентной (!) доли ошибок, тогда как уже двадцатипроцентная доля артефактов при обычном АБД ряд авторов по данной тематике считает недопустимой.

Заключение

Таким образом, обсуждаемый подход позволяет:

1. Выполнять анализ импликативных связей на основе нестрогих вероятностей при использовании как качественных, так и количественных данных, в том числе, в условиях их низкой достоверности и противоречивости.

2. Учитывать степень доверия к данным. Последнее особенно важно, если в данных содержатся артефакты, что характерно для АБД.

3. Учитывать влияние артефактов, а также малодостоверных и противоречивых данных на результат индуктивного вывода при значительной доле таких данных, что существенно при работе с соответствующими массивами в ходе АБД.

Литература

1. Формула Big Data: семь «V» + неординарная задача. URL: https://www.fsight.ru/blog/formula-big-data-sem-v-neordinarnaja-zadacha-2/
2. Лобанов А.А. Большие данные: проблемы обработки // Вестник МГТУ МИРЭА. 2014. № 3(4). С. 51-58.
3. Абрамова А.А. Анализ использования больших данных для принятия решений в промышленной сфере // Экономика и качество систем связи. 2023, № 3. С. 13-21.
4. Кельчевская Н.Р., Колясников М.С. Использование больших данных в стратегическом управлении знаниями компании, следующей трендам Индустрии 4.0 // Лидерство и менеджмент. 2020. Том 7. № 3. С. 405-426. doi: 10.18334/lim.7.3.110662.
5. Fosso Wamba S. et al. How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study // International Journal of Production Economics.

2015. Vol. 165. pp. 234-246. doi: 10.1016/j.ijpe.2014.12.031.
6. Орешков В.И. Интеллектуальный анализ данных как современный инструмент поддержки управленческих решений // Вестник Рязанского государственного агротехнологического университета имени П.А. Костычева. 2011. № 4 (12). С. 55-59.
7. Емельченков Е.П. Большие данные. Методы интеллектуального анализа // Системы компьютерной математики и их приложения. 2013. № 14. С. 75-79.
8. Есауленко А.С., Никоненко Н.Д. Большие данные. Реальность и перспективы // Управление инновациями: теория, методология, практика. 2016. № 17. С. 74-79.
9. Медведев Д.А. Большие данные: причины появления и как их можно использовать // Наука и образование сегодня. 2019. № 4(39). С. 14-16.
10. Кузора С.С., Натаров И.П. Цифровая трансформация и большие данные // Вестник Российского университета дружбы народов. Серия: Государственное и муниципальное управление. 2022. Т. 9. № 2. С. 150–161. doi: 10.22363/2312-8313-2022-9-2-150-161.
11. Магеррамов З.Т., Абдуллаев В.Г., Магеррамова А.З. Big Data: проблемы, методы анализа, алгоритмы // Радиоэлектроника и информатика, 2017. №3. С. 42-52.
12. Арустамов А. Критерии качества данных, 2023. URL: https://loginom.ru/blog/data-quality-criteria.
13. Дударев В.А. Подход к заполнению пропусков в обучающих выборках для компьютерного конструирования неорганических соединений // Вестник МИТХТ. 2014, Т.9, №1. С. 73-75.
14. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, 2004, №3. С. 1-20.
15. Финн В. К. Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез // Искусственный интеллект и принятие решений, 2010. №4. С. 41-48.
16. Виноградов Д.В. Анализ результатов применения ВКФ-системы: успехи и открытая проблема // Научно-техническая информация. Серия 2: Информационные процессы и системы, 2017. № 5. С. 1-4.
17. Панов А. И. Выявление причинно-следственных связей в данных психологического тестирования логическими методами // Искусственный интеллект и принятие решений, 2013. №1. С. 24–32.
18. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / Сост. О.М. Аншаков, Е.Ф. Фабрикантова. М.: Книжный дом «ЛИБРИКОМ», 2009. – 432 с.
19. Dunn J.M. Algebra of Intensional Logics. Doctoral Dissertation University of Pittsburg, Ann Arbor, 1966.
20. Dunn J.M. Intuitive semantics for first-degree entailment and “coupled trees” // Philosophical Studies.Vol. 29, 1976. – pp.149-158.
21. Belnap N. A useful four-valued logic // J.M.Dunn and G.Epstein (eds.). Modern Uses of Multiple-Valued Logic. – Dordrecht: D. Reidel Publish. Co., 1977. – pp. 8-37.
22. Belnap N. How a computer should think // G. Ryle (ed.). Contemporary Aspects of Philosophy. – Stocksfield: Oriel Press Ltd., 1977. – P. 30-55.
23. Аршинский Л.В. Методы обработки нестрогих высказываний. Иркутск: изд-во ВСИ МВД России, 1998. 40 с.
24. Ивлев Ю.В. Логика: Учебник 3-е изд. М.: ТК Велби, Изд-во Проспект, 2004. 288 с.
25. Голенков В.В. Статистические основы индуктивного вывода: учеб. пособие. Минск: БГУИР, 2009. 202 с.
26. Кайберг Г. Вероятность и индуктивная логика / Пер. с англ. – М.: Изд-во «Прогресс», 1978. 373 с. (Kyburg H.E. Probability and Inductive Logic. – L.: Macmillan, 1970. 272 p.).
27. Inductive Inference. URL: https://www.sciencedirect.com/topics/mathematics/inductive-inference
28. Аршинский Л.В., Лебедев В.С. Объективизация баз знаний интеллектуальных систем на основе индуктивного вывода с использованием нестрогих вероятностей // Информационные и математические технологии в науке и управлении. 2022. № 4(28). С. 190-200. doi:10.38028/ESI.2022.28.4.015.
29. Аршинский Л.В. Приложение логик с векторной семантикой к описанию случайных событий и оценке риска // Проблемы анализа риска. 2005. Т.2. № 3. С.231-248.
30. Нечеткая логика в моделях управления и искусственного интеллекта / под ред. Д.А. Поспелова. М.: Наука. Гл. ред. физ.-мат. лит., 1986. 312 с.
31. Gottwald S. Treatise on Many-Valued Logics. Leipzig, 2000. 604 p.
32. Аршинский Л.В. Оценка истинности взаимоисключающих гипотез средствами векторной логики // Информационные и математические технологии/ Труды Байкальской Всероссийской конференции «Информационные и математические технологии». Иркутск: ИСЭМ СО РАН, 2004. С. 188-194.
33. Pima Indians Diabetes – EDA & Prediction (0.906). URL: https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/input.
34. Уровень сахара в крови: норма, установленная ВОЗ для здоровых людей. URL: https://yandex.ru/health/turbo/articles?id=4419.

Аршинский Леонид Вадимович. Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения», Иркутск, Россия. Профессор кафедры «Информационные системы и защита информации», доктор технических наук, доцент. Область научных интересов: системный анализ, информационные технологии, искусственный интеллект, информационная безопасность, гидроаэродинамика. E-mail: larsh@mail.ru

Вадим Сергеевич Лебедев. Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения», Иркутск, Россия. Аспирант.Область научных интересов: информационные технологии, искусственный интеллект. E-mail: lebedevvs97@yandex.ru

Processing of Data for Inductive Inference Based on Non-Strict Probability

L. V. Arshinskiy, V. S. Lebedev

Irkutsk State Transport University, Irkutsk, Russia

Abstract. Based on methods of inductive logic, an approach to identifying of implication relationships "If A, then b" in Big Data is considered. This approach is considered in conditions of low reliability and inconsistency of data. To work in this condition, logics with vector semantics in the form of V^{TF} logics are used. The presence or absence of phenomena in tables of their joint occurrence is formalized by truth vectors with components v^+ and v^-, where v^+ is a measure of the true of a statement about the presence of a phenomenon, v^- is a measure of its false. On the base of statistical induction principal, the indicator of the validity of a causal relationship is calculated as the average value of the truth vectors of the corresponding non-strict propositions. The resulting value is interpreted as a non-strict probability of the relationship, which acts as a vector indicator of its validity. The applicability of the approach for processing qualitative and quantitative data, as well as data containing artifacts, is shown.

Keywords: big data, data mining, inductive inference, non-strict probability, logic with vector semantics.

DOI 10.14357/20718632240201 **EDN** HUCOJV

References

1. Formula Big Data: sem` «V» + neordinarnaya zadacha [Big Data formula: seven "Vs" + an extraordinary task]. Available at: https://www.fsight.ru/blog/formula-big-data-sem-v-neordinarnaja-zadacha-2/ (accessed January 10, 2024)
2. Lobanov, A.A. 2014. Bol'shie dannye: problemy obrabotki [Big data: processing problems]. Vestnik MGTU MIREA [Bulletin of MSTU MIREA]. 3:51-58.
3. Abramova, A.A. 2023. Analiz ispol'zovaniya bol'shih dannyh dlya prinyatiya reshenij v promyshlennoj sfere [Analysis of the use of big data for decision making in the industrial sector]. Ekonomika i kachestvo sistem svyazi [Economics and quality of communication systems]. 3:13-21.
4. Kel'chevskaya, N.R., and M.S. Kolyasnikov. 2020. Ispol'zovanie bol'shih dannyh v strategicheskom upravlenii znaniyami kompanii, sleduyushchej trendam Industrii 4.0 [The use of big data in the strategic knowledge management of a company following the trends of Industry 4.0]. Liderstvo i menedzhment [Leadership and Management]. 7(3):405-426. doi: 10.18334/lim.7.3.110662.
5. Fosso Wamba, S. et al. 2015. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics. 165: 234-246. doi: 10.1016/j.ijpe.2014.12.031.
6. Oreshkov, V.I. 2011. Intellektual'nyj analiz dannyh kak sovremennyj instrument podderzhki upravlencheskih reshenij [Data mining as a modern tool for supporting management decisions]. Vestnik Ryazanskogo gosudarstvennogo agrotekhnologicheskogo universiteta [Bulletin of the Ryazan State Agrotechnological University]. 4:55-59.
7. Emel'chenkov, E.P. 2013. Bol'shie dannye. Metody intellektual'nogo analiza [Big Data. Methods of intellectual analysis]. Sistemy komp'yuternoj matematiki i ih prilozheniya [Systems of computer mathematics and their applications]. 14:75-79.
8. Esaulenko, A.S. and N.D. Nikonenko. 2016. Bol'shie dannye. Real'nost' i perspektivy [Big data. Reality and prospects]. Upravlenie innovaciyami: teoriya, metodologiya, praktika [Innovation management: theory, methodology, practice]. 17:74-79.
9. Medvedev, D.A. 2019. Bol'shie dannye: prichiny poyavleniya i kak ih mozhno ispol'zovat' [Big data: reasons for its appearance and how it can be used]. Nauka i obrazovanie segodnya [Science and Education Today]. 4:14-16.
10. Kuzora, S.S. and I.P. Natarov. 2022. Cifrovaya transformaciya i bol'shie dannye [Digital transformation and big data]. Vestnik Rossijskogo universiteta druzhby narodov. Seriya: Gosudarstvennoe i municipal'noe upravlenie [Bulletin of the Russian Peoples' Friendship University. Series: State and municipal administration]. 9(2):150–161. doi: 10.22363/2312-8313-2022-9-2-150-161.
11. Magerramov, Z.T., V.G. Abdullaev and A.Z. Magerramova. 2017. Big Data: problemy, metody analiza, algoritmy [Big Data: problems, analysis methods, algorithms]. Radioelektronika i informatika [Radioelectronics and Informatics]. 3:42-52.
12. Kriterii kachestva dannyh [Data quality criteria]. Available at: https://loginom.ru/blog/data-quality-criteria (accessed at 10 January, 2024).
13. Dudarev, V.A. 2014. Podhod k zapolneniyu propuskov v obuchayushchih vyborkah dlya komp'yuternogo konstruirovaniya neorganicheskih soedinenij [An approach to filling gaps in training samples for computer-aided design of inorganic compounds]. Vestnik MITHT [Bulletin of MITHT]. 9(1):73-75.
14. Finn, V.K. 2004. Ob intellektual'nom analize dannykh [On intelligent data analysis]. Novosti iskusstvennogo intellekta [Artificial Intelligence News]. 3:1-20.

15. Finn, V. K. 2010. Ob opredelenii empiricheskikh zakonomernostey posredstvom DSM - metoda avtomaticheskogo porozhdeniya gipotez [On the determination of empirical patterns using JSM - the method of automatic generation of hypotheses]. Iskusstvennyy intellekt i prinyatiye resheniy [Artificial intelligence and decision making].4:41-48.
16. Vinogradov, D.V. 2017. Analiz rezul'tatov primeneniya VKF-sistemy: uspekhi i otkrytaya problema [Analysis of the results of using the VKF system: successes and an open problem]. Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnyye protsessy i sistemy [Scientific and technical information. Series 2: Information processes and systems]. 5:1-4.
17. Panov, A.I. 2013. Vyyavleniye prichinno-sledstvennykh svyazey v dannykh psikhologicheskogo testirovaniya logicheskimi metodami [Identification of cause-and-effect relationships in psychological testing data using logical methods]. Iskusstvennyy intellekt i prinyatiye resheniy [Artificial intelligence and decision making]. 1:24–32.
18. Anshakov, O.M. et al. 2009. DSM-metod avtomaticheskogo porozhdeniya gipotez: Logicheskiye i epistemologicheskiye osnovaniya [JSM method for automatically generating hypotheses: Logical and epistemological foundations]. Moscow. Book house “LIBRIKOM”. 432 p.
19. Dunn, J.M. 1966. Algebra of Intensional Logics. Doctoral Dissertation University of Pittsburg, Ann Arbor.
20. Dunn, J.M. 1976. Intuitive semantics for first-degree entailment and “coupled trees”. Philosophical Studies. 29:149-158.
21. Belnap, N. 1977. A useful four-valued logic. Modern Uses of Multiple-Valued Logic. Dordrecht: D. Reidel Publish. Co. 8-37.
22. Belnap N. 1977. How a computer should think. Contemporary Aspects of Philosophy. Stocksfield: Oriel Press Ltd. 30-55.
23. Arshinskiy, L.V. eds. 1998. Metody obrabotki nestrogih vyskazyvanij [Methods for processing non-strict proposition]. Irkutsk: East-Siberian Institute of MIA of Russia. 40 p.
24. Ivlev, Yu.V. eds. 2004. Logika: Uchebnik 3-e izd [Logic: Textbook 3rd ed.]. Moscow: TK Welby, Prospekt Publishing House. 288 p.
25. Golenkov, V.V. eds. 2009. Statisticheskie osnovy induktivnogo vyvoda: ucheb. posobie [Statistical foundations of inductive inference: textbook]. Minsk: BSUIR. 202 p.
26. Kyburg, H.E. 1970. Probability and Inductive Logic. L.: Macmillan. 272 p.
27. Inductive Inference. Available at: https://www.sciencedirect.com/topics/mathematics/ inductive-inference (accessed at 10 January, 2024).
28. Arshinskiy, L.V. and V.S. Lebedev. 2022. Ob"ektivizaciya baz znanij intellektual'nyh sistem na osnove induktivnogo vyvoda s ispol'zovaniem nestrogih veroyatnostej [Objectification of intelligent systems knowledge bases based on the inductive inference using non-strict probabilities]. Informacionnye i matematicheskie tekhnologii v nauke i upravlenii [Information and mathematical technologies in science and management]. 4:190-200. doi:10.38028/ESI.2022.28.4.015.
29. Arshinskiy L.V. 2005. Prilozhenie logik s vektornoj semantikoj k opisaniyu sluchajnyh sobytij i ocenke riska [Application of vector semantics logics for description of occasion events and risks evaluation] // Problemy analiza riska [Issues of risk analysis]. 2(3):231-248.
30. Nechetkaya logika v modelyah upravleniya i iskusstvennogo intellekta / pod red. D.A. Pospelova [Fuzzy logic in control models and artificial intelligence / ed. YES. Pospelov], eds. 1986. M.: Science. Ch. ed. physics and mathematics lit. 312 p.
31. Gottwald, S. 2000. Treatise on Many-Valued Logics. Leipzig. 604 p.
32. Arshinskiy, L.V. 2004. Ocenka istinnosti vzaimoisklyuchayushchih gipotez sredstvami vektornoj logiki [Assessing the truth of mutually exclusive hypotheses using vector logic]. Informacionnye i matematicheskie tekhnologii / Trudy Bajkal'skoj Vserossijskoj konferencii «Informacionnye i matematicheskie tekhnologii» [Information and mathematical technologies/ Proceedings of the Baikal All-Russian conference “Information and mathematical technologies”]. Irkutsk. 188-194.
33. Pima Indians Diabetes – EDA & Prediction (0.906). URL: https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906/input.
34. Uroven' sahara v krovi: norma, ustanovlennaya VOZ dlya zdorovyh lyudej [Blood sugar level: the norm established by WHO for healthy people]. Available at: https://yandex.ru/health/turbo/articles?id=4419 (accessed at 10 January, 2024).

Arshinskiy Leonid V. Professor, Irkutsk State Transport University, 15 Chernyshevskogo str., 15. Irkutsk, 664074, Russia, e-mail: larsh@mail.ru

Lebedev Vadim S. Graduate student, Irkutsk State Transport University, 15 Chernyshevskogo str., 15. Irkutsk, 664074, Russia, e-mail: lebedevvs97@yandex.ru