УДК 004.89+510.644

Л.В. Аршинский [1], В.С. Лебедев [1]

[1] Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

КОНКУРС ГИПОТЕЗ ПРИ ИНДУКТИВНОМ ВЫВОДЕ НА ОСНОВЕ НЕСТРОГИХ ВЕРОЯТНОСТЕЙ

Аннотация. В работе описана методика выбора гипотез вида «Если…, то…» из нескольких альтернатив, полученных с помощью объединённого метода сходства и различия Бэкона-Милля, в условиях, когда соответствующая информация является малодостоверной и/или противоречивой. Источниками информации могут быть хорошо структурированные массивы: электронные таблицы, реляционные базы данных и т.п. В основе методики лежит понятие нестрогой вероятности, вытекающее из векторного представления истинности для V^{TF}-логик. Выбор конкретной гипотезы предлагается осуществлять с помощью мер определённости и достоверности, известных в V^{TF}-логиках и применяемых также для нестрогих вероятностей.

Ключевые слова: анализ данных, индуктивный вывод, объединённый метод сходства и различия, нестрогая вероятность.

L.V. Arshinskiy [1], V.S. Lebedev [1]

[1] Irkutsk State Transport University, Irkutsk, Russian Federation

HYPOTHESIS CONTEST FOR INDUCTIVE INFERENCE BASED ON NONSTRICT PROBABILITIES

Abstract. The paper describes a method for selecting hypotheses of the form "If ..., then..." from several alternatives obtained using the Bacon-Mill joint method of agreement and difference, in conditions when the relevant information is unreliable and/or contradictory. Sources of information can be well-structured arrays: spreadsheets, relational databases, etc. The methodology is based on the concept of non-strict probability, which follows from the vector representation of truth for V^{TF}-logics. The choice of a specific hypothesis is proposed to be carried out using measures of certainty and reliability, known in V^{TF}-logics and also used for non-strict probabilities.

Keywords: data mining, inductive inference, joint method of agreement and difference, non-strict probability.

Введение. В работе авторов [1] рассмотрено применение нестрогих вероятностей в задаче индуктивного поиска скрытых закономерностей в таблицах данных. Рассматривались простейшие закономерности вида:

$$a_i \rightarrow b \qquad (1)$$

как результат анализа таблиц совместной встречаемости явлений на основе объединённого метода сходства и различия (рис. 1).

	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

Рис. 1. Пример таблицы совместной встречаемости явлений a и b.

Здесь a_{ik} и b_k равны 1 (явление наблюдается) или 0 (явление не наблюдается).

Источником таких таблиц могут выступать реляционные базы данных, электронные таблицы, лабораторные журналы и т.д. Главная особенность такого подхода – возможность работать с неполной и противоречивой информацией о наблюдении явлений a и b.

Идеальный случай применения метода сходства и различия предусматривает:

1) достоверно наблюдаемые факты a и b;

2) ситуацию, когда каждая строка таблицы совместной встречаемости подтверждает зависимость (1), если она имеется.

Однако в реальности вполне возможно, что:

1) зависимость (1) подтверждается не всеми строками;
2) факты *a* и *b* могут быть не наблюдаемыми с достоверностью.

При этом отсутствие достоверной наблюдаемости может заключаться в:

1) противоречивости сведений об их наблюдении;
2) отсутствии или низкой надёжности сведений об их наличии/отсутствии.

Причем обе эти ситуации могут реализовываться совместно.

Первая ситуация разрешается популярным в индуктивной логике вероятностно-статистическим подходом, когда учитывается относительная доля строк, подтверждающих (1) (см. напр. [2-5]). Это строки, в которых для конкретного a_i истинны конъюнкции $a_i\&b$ и $\neg a_i\&\neg b$, т.е. истинна дизъюнкция:

$$a_i\&b\vee\neg a_i\&\neg b. \qquad (2)$$

В этом случае говорят о «вероятности», «степени правдоподобия» соответствующего суждения. В рассматриваемом случае – суждения (1).

Вторая проблема сложнее. Традиционный индуктивный вывод не рассматривает, вообще говоря, ситуации, когда степень неуверенности в факте столь высока, что сложно выбрать что поставить в соответствующую ячейку таблицы совместной встречаемости: 0 или 1. Да и подстановка конкретно нуля или единицы в условиях неопределённости выбора – это «заметание сора под ковёр», уход от факта незнания такого значения. Именно для таких случаев и был предложен подход, описанный в [1]. Сам подход опирается на понятие нестрогой вероятности, введённое в [6,7].

Нестрогая вероятность. В [6] понятие нестрогой вероятности определялось как вероятность события, относительно составляющих которого нет полной уверенности, что они входят в соответствующее событию подмножество *A* полной группы событий Ω. Такая ситуация возможна, когда решение о принадлежности/непринадлежности элементарного события ω подмножеству *A* принимается на основе свидетельств, которые могут противоречить друг другу и/или их источники не заслуживают доверия. Для таких случаев в [6] рассматривалась истинность предиката $F(\omega,A)$ = «Элементарное событие ω благоприятно с точки зрения события *A*». Если, как это принято, представить вероятность события *A* суммой:

$$p(A)=\sum_{\omega\in A}p(\omega), \qquad (3)$$

то такой же результат даст сумма:

$$p(A)=\sum_{\omega\in\Omega}\|F(\omega,A)\|\,p(\omega), \qquad (4)$$

где $\|F(\omega,A)\|$ – истинность $F(\omega,A)$. Истинность равна 1, если $\omega\in A$, и 0 в противном случае. Векторное представление истинности предиката:

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

где $F^{+}(\omega,A)$ – совокупный вклад свидетельств в пользу $F(\omega,A)$, а $F^{-}(\omega,A)$ – совокупный вклад опровергающих свидетельств, порождает иное, векторное представление вероятности:

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

Оно отражает факт отсутствия твердой уверенности в благоприятности/неблагоприятности ω для *A* (имеются доводы «за» и «против» с разной степенью доверия к ним).

Очевидно, что для строгих значений вектора $\|F(\omega,A)\|$, равных $\langle 1;0\rangle$ или $\langle 0;1\rangle$, (3) превращается в (4), а (4) в привычную вероятность, где $P^{+}(A)$ – вероятность *A*, а $P^{-}(A)$ – вероятность противоположного события. В [6] для такого представления даны выражения для сложных вероятностей: отрицания, суммы и произведения (в двух формах каждая). Для индуктивного вывода хорошо подходят первые формы:

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

– вероятность первой формы противоположного события;

$$P(A \vee B) = \langle \sum_{\omega\in\Omega} [F^+(\omega, A) \oplus F^+(\omega, B)]\, p(\omega);\ \sum_{\omega\in\Omega} [F^-(\omega, A) \bullet F^-(\omega, B)]\, p(\omega) \rangle$$

– вероятность первой формы суммы двух нестрогих событий;

$$P(A \,\&\, B) = \langle \sum_{\omega\in\Omega} [F^+(\omega, A) \bullet F^+(\omega, B)]\, p(\omega);\ \sum_{\omega\in\Omega} [F^-(\omega, A) \oplus F^-(\omega, B)]\, p(\omega) \rangle$$

– вероятность первой формы произведения двух нестрогих событий;

Здесь $\bullet$ и $\oplus$ – соответственно триангулированная (треугольная) норма и ко-норма в инфиксной записи, с дополнительной аксиомой:

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

(или, что то же самое: $(1-x) \oplus (1-y) = 1 - x \bullet y$). Двумя наиболее распространёнными примерами данных норм выступают известные пары функций:

$$x \bullet y = x \cdot y; \qquad x \oplus y = x + y - x \cdot y;$$
$$x \bullet y = \min(x, y); \qquad x \oplus y = \max(x, y).$$

Смысл первых форм сложных событий достаточно очевиден. Для противоположного события благоприятность и неблагоприятность меняются местами. Для суммы достаточно, чтобы элементарное событие ω было благоприятным было хотя бы для одного из подмножеств A или B и неблагоприятным для обоих сразу. Для произведения – чтобы благоприятным для обоих и неблагоприятным хотя бы для одного.

В [6] нестрогая вероятность интерпретировалась как риск («нестрогий риск», когда соответствующие случайные события оцениваются как с вредной, так и полезной стороны), возможны и другие интерпретации. Одну из интерпретаций предлагает индуктивный вывод.

Нестрогая вероятность в индуктивном выводе. В [1] понятие нестрогой вероятности используется для оценки достоверности импликации (1) по результату анализа таблицы совместной встречаемости. Для этого каждая строка таблицы рассматривается как аргумент за или против импликации. В классическом случае строка свидетельствует в пользу (1), если истинно высказывание (2). Это так, когда a_i и b совместно наблюдаются или не наблюдаются – значения 1 и 1, либо 0 и 0 в соответствующих позициях строки (объединённый метод сходства и различия). Согласно вероятностно-статистическому взгляду на индукцию, доля строк для которых истинность $\|a_i \& b \vee \neg a_i \& \neg b\|$=1 показывает степень справедливости суждения (1). Это можно интерпретировать как «вероятность» гипотезы (1) или как её (нечёткую) истинность.

Если свидетельства в пользу a_i и b неочевидны (т.е. данные о них малодостоверны и/или противоречивы) истинность каждого факта в таблице представляется вектором $\langle t^+; t^- \rangle$, отражающим степени уверенности в его наличии (t^+) или отсутствии (t^-) (рис. 2).

	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	$\langle a_{1K}^+; a_{1K}^- \rangle$	$\langle a_{2K}^+; a_{2K}^- \rangle$	…	$\langle a_{nK}^+; a_{nK}^- \rangle$	$\langle b_K^+; b_K^- \rangle$

Рис. 2. Пример таблицы совместной встречаемости явлений a и b для малодостоверной и/или противоречивой информации

Аргументом в пользу гипотезы (1) для каждой строки здесь также выступает истинность высказывания (2), но в векторном варианте. Соответственно вклад каждой строки рассчитывается как:

$$\| F(k, a_i \to b) \| = \| a_{ik} \,\&\, b_k \vee \neg a_{ik} \,\&\, \neg b_k \| . \qquad (5)$$

Присваивая каждой строке «вероятность» $p(k)=1/K$ и принимая, что векторная истинность конъюнкции, дизъюнкции и отрицания определяются их первыми формами [6]:

$$\| A \,\&\, B \| = \langle A^+ \bullet B^+ ; A^- \oplus B^- \rangle ;$$

$$\| A \vee B \| = \langle A^+ \oplus B^+ ; A^- \bullet B^- \rangle ;$$

$$\| \neg A \| = \langle A^- ; A^+ \rangle ;$$

степень уверенности в (1) рассчитывается согласно (4) как нестрогую вероятность:

$$P(a_i \to b) = \langle \frac{1}{K} \sum_{k=1}^{K} [a_{ik}^+ \bullet b_k^+ \oplus a_{ik}^- \bullet b_k^-] ; \frac{1}{K} \sum_{k=1}^{K} [(a_{ik}^- \oplus b_k^-) \bullet (a_{ik}^+ \oplus b_k^+)] \rangle . \qquad (6)$$

Именно этот показатель и будет характеризовать уверенность в гипотезе (1) при дефиците и противоречивости информации об a_i и b.

Конкурс гипотез. Рассматриваемая техника представлена для выделенной гипотезы, относительно которой делается заключение о степени её достоверности (истинности). Однако возможна ситуация, когда заранее неизвестно какой из фактов a_i является или может являться причиной b. Возникает конкурс гипотез, из которых нужно выбрать лучшую. Это можно сделать, исследовав всё множество потенциальных причин $\{a_i\}$. Иначе говоря, на основании таблицы совместной встречаемости определить какая из гипотез

$$a_1 \to b;$$
$$a_2 \to b;$$
$$\dots$$
$$a_n \to b$$

предпочтительнее. Если утверждения о наблюдениях a_{ik} и b_k строго истинные или ложные, т.е. в ячейках таблицы находятся только 0 или 1, выбор можно сделать, рассчитав «вероятность» каждой гипотезы с помощью (4). Предпочтение естественно отдать той, для которой значение

$$p_i = \frac{1}{K} \sum_{k=1}^{K} \| F(k, a_i \to b) \|$$

больше; p_i – «вероятность», что гипотеза $a_i \to b$ справедлива, как доля строк, в которых (5) принимает значение 1.

В нестрогом случае, когда в ячейках таблицы стоят вектора $\langle a_{ik}^+ ; a_{ik}^- \rangle$, значения вероятностей $P(a_i \to b)$ линейно не упорядочены. Для выбора предпочтительной гипотезы следует использовать дополнительные показатели [7]. К таким показателям целесообразно отнести меру определённости:

$$\mu_{\text{о}} (A) = P(A^+) \oplus P(A^-);$$

и мера достоверности:

$$\mu_{\text{д}} (A) = P(A^+) - P(A^-).$$

Первая позволяет исключать из рассмотрения гипотезы, не подтверждённые убедительными свидетельствами за или против. Вторая – выбрать наиболее предпочтительные гипотезы с точки зрения соотношения свидетельств.

Для нестрогих вероятностей известны также мера строгости в форме:

$$\mu_{\text{с}} (A) = P(A^+) \oplus P(A^-) - P(A^+) \bullet P(A^-),$$

или

$$\mu_{\text{с}} (A) = |P(A^+) - P(A^-)|;$$

мера противоречивости:

$$\mu_{\text{п}} (A) = P(A^+) \bullet P(A^-);$$

и мера избыточности:

$$\mu_{\text{изб}} (A) = P(A^+) + P(A^-) - 1;$$

однако первые две выглядят наиболее подходящими.

В завершение рассмотрим небольшой модельный пример выбора предпочтительной гипотезы из трёх возможных. Числовые значения в таблице произвольны.

Таблица 1.

Совместная встречаемость явлений a_1, a_2, a_3 и b.

	a_1	a_2	a_3	b
1	⟨0.6;0.2⟩	⟨0.7;1⟩	⟨1;0.8⟩	⟨0.9;0⟩
2	⟨0;1⟩	⟨0.3;0.9⟩	⟨0.1;0.1⟩	⟨0.1;0.8⟩
3	⟨0.8;0.1⟩	⟨0.5;0.7⟩	⟨0.4;0.2⟩	⟨1;0.2⟩
4	⟨0.9;0.3⟩	⟨0.9;0.5⟩	⟨0.3;0.2⟩	⟨0.9;0.3⟩

Мера определённость для гипотез приняла значение:

$$\mu_{\text{о}}(a_1 \to b) = 0.81;$$
$$\mu_{\text{о}}(a_2 \to b) = 0.84;$$
$$\mu_{\text{о}}(a_3 \to b) = 0.48.$$

Задавая порог определённости 0.5, третью гипотезу по этому показателю исключаем как малообоснованную. По мере достоверности получаем:

$$\mu_{\text{д}}(a_1 \to b) = 0.47;$$
$$\mu_{\text{д}}(a_2 \to b) = 0.22.$$

Наибольшая достоверность у первой гипотезы, её и следует принять. Нормы и ко-нормы здесь рассчитывались как $x \cdot y$ и $x + y - x \cdot y$ соответственно.

Заключение. Представленная методика позволяет автоматизировать поиск и отбор гипотез о причинно-следственных связях, неявно присутствующих в реляционных базах данных, электронных таблицах и иных аналогичным образом структурированных массивах в условиях малодостоверной и противоречивой информации. Если эти проблемы отсутствуют, она естественным образом превращается в методику построения гипотез на основе объединённого метода сходства и различия в духе Бэкона-Милля с учётом доли подтверждающих и опровергающих свидетельств.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Аршинский Л.В., Лебедев В.С. Применение нестрогой вероятности в задачах индуктивного вывода // Актуальные вопросы прикладной дискретной математики. Сборник научных трудов. Сер. «Дискретный анализ и информатика». – Иркутск, 2022. – С. 9-16.

2. Голенков В.В. Статистические основы индуктивного вывода: учеб. пособие / В.В. Голенков, М. Д. Степанова, С.А. Самодумкин, Н.А. Гулякина. – Минск: БГУИР, 2009. – 202 с.

3. Кайберг Г. Вероятность и индуктивная логика / Г. Кайберг. – М.: Изд-во «Прогресс», 1978. – 373 с.

4. Inductive Logic // Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/ entries/logic-inductive.

5. Inductive Inference // ScienceDirect. URL: https://www.sciencedirect.com/topics/ mathematics/inductive-inference.

6. Аршинский Л.В. Приложение логик с векторной семантикой к описанию случайных событий и оценке риска / Л.В. Аршинский. // Проблемы анализа риска, 2005. –Т.2. – № 3. – С.231-248.

7. Аршинский Л.В. Векторные логики: основания, концепции, модели / Л.В. Аршинский. – Иркутск: Иркут. гос. ун-т, 2007. – 228 с

REFERENCES

1. Arshinsky L.V., Lebedev V.S. *Primenenie nestrogoj veroyatnosti v zadachah induktivnogo vyvoda* [Application of non-strict probability in problems of inductive inference] // *Aktual'nye voprosy prikladnoj diskretnoj matematiki. Sbornik nauchnyh trudov. Ser. «Diskretnyj analiz i in-*

formatika» [Actual questions of applied discrete mathematics. Collection of scientific papers. Ser. "Discrete analysis and computer science".]. Irkutsk, 2022. pp. 9-16. (in Russian).

2. Golenkov V.V., Stepanova M.D., Samodumkin S.A., Gulyakina N.A. Statisticheskie osnovy induktivnogo vyvoda: ucheb. posobie [Statistical bases of inductive inference: textbook]. Minsk: BGUIR, 2009, 202 p. (in Russian).

3. Kyburg H.E. Veroyatnost' i induktivnaya logika [Probability and Inductive Logic]. Moscow: Izd-vo «Progress», 1978, 373 p. (in Russian).

4. Inductive Logic // Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/ entries/logic-inductive.

5. Inductive Inference // ScienceDirect. URL: https://www.sciencedirect.com/topics/ mathematics/inductive-inference.

6. Arshinskiy L.V. Prilozhenie logik s vektornoj semantikoj k opisaniyu sluchajnyh sobytij i ocenke riska [Application of logic with vector semantics to the description of random events and risk assessment] // Problemy analiza riska [Issues of Risk Analysis], 2005, vol.2, no. 3, pp. 231-248. (in Russian)/

7. Arshinskiy L.V. Vektornye logiki: osnovanija, koncepcii, modeli [Vector logic: foundations, concepts, models]. Irkutsk: Irkutskij gosudarstvennyj universitet [Irkutsk state university], 2007, 228 p. (in Russian)

Информация об авторах

Леонид Вадимович Аршинский – д. т. н., доцент, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: larsh@mail.ru.

Вадим Сергеевич Лебедев – аспирант кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: lebedevvs97@yandex.ru.

Authors

Leonid Vadimovich Arshinskiy – Doctor of Technical Science, professor of department “Information Systems and Information Security”, Irkutsk State Transport University, Irkutsk, e-mail: larsh@mail.ru.

Vadim Sergeevich Lebedev – postgraduate student of department “Information Systems and Information Security”, Irkutsk State Transport University, Irkutsk, e-mail: lebedevvs97@yandex.ru.

Для цитирования

Аршинский Л.В., Лебедев В.С. Конкурс гипотез при индуктивном выводе на основе нестрогих вероятностей // Информационные технологии и математическое моделирование в управлении сложными системами: электрон. науч. журн. – 2022. – №4(16). – С. 10-15 – DOI: 10.26731/2658-3704.2022.4(16).10-15 – Режим доступа: http://ismm-irgups.ru/toma/416-2022, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 17.12.2022).

For citations

Arshinskiy L.V., Lebedev V.S Hypothesis contest for inductive inference based on non-strict probabilities // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2022. No. 4(16). P. 10-15. DOI: 10.26731/2658-3704.2022.4(16).10-15 [Accessed 17/12/22].